UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE ESTUDIOS SUPERIORES CUAUTITLÁN

Estudio de sesgos o prejuicios en el aprendizaje computacional: enfoque en el diseño de modelos con clasificadores bayesianos

T E S I S

QUE PARA OBTENER EL TÍTULO DE

INGENIERO EN TELECOMUNICACIONES, SISTEMAS Y ELECTRÓNICA

P R E S E N T A

MAURICIO BYRD VICTORICA

Asesor: Ing. José Luis Barbosa Pacheco

Cuautitlán Izcalli, Estado de México 2021

VNIVERSIDAD NACIONAL AVTNOMA DE MEXICO

FACULTAD DE ESTUDIOS SUPERIORES CUAUTITLÁN
SECRETARÍA GENERAL
DEPARTAMENTO DE EXÁMENES PROFESIONALES

U. N. A. M.
FACULTAD DE ESTUDIOS SUPERIORES-CUAUTITLÁN

ASUNTO: VOTO APROBATORIO

M. en C. JORGE ALFREDO CUÉLLAR ORDAZ
DIRECTOR DE LA FES CUAUTITLAN
PRESENTE

ATN: I.A. LAURA MARGARITA CORTAZAR FIGUEROA
Jefa del Departamento de Exámenes Profesionales
de la FES Cuautitlán.

Con base en el Reglamento General de Exámenes, y la Dirección de la Facultad, nos permitimos comunicar a usted que revisamos el: **Trabajo de Tesis**

Estudio de sesgos o prejuicios en el aprendizaje computacional: enfoque en el diseño de modelos con clasificadores bayesianos.

Que presenta el pasante: **Mauricio Byrd Victorica**
Con número de cuenta: **312530336** para obtener el título de: **Ingeniero en Telecomunicaciones, Sistemas y Electrónica**

Considerando que dicho trabajo reúne los requisitos necesarios para ser discutido en el **EXAMEN PROFESIONAL** correspondiente, otorgamos nuestro **VOTO APROBATORIO.**

ATENTAMENTE
"POR MI RAZA HABLARÁ EL ESPÍRITU"
Cuautitlán Izcalli, Méx. a 23 de Septiembre de 2020.

PROFESORES QUE INTEGRAN EL JURADO

	NOMBRE	FIRMA
PRESIDENTE	Mtro. Leopoldo Martín del Campo Ramírez	
VOCAL	Ing. José Luis Barbosa Pacheco	
SECRETARIO	Mtro. José Isaac Sánchez Guerra	
1er. SUPLENTE	Lic. Mauricio Jaques Soto	
2do. SUPLENTE	Dr. David Tinoco Varela	

NOTA: los sinodales suplentes están obligados a presentarse el día y hora del Examen Profesional (art. 127).

LMCF/cga*

VNIVERSIDAD NACIONAL AVPNOMA DE MEXICO

FACULTAD DE ESTUDIOS SUPERIORES CUAUTITLÁN
SECRETARÍA GENERAL
DEPARTAMENTO DE EXÁMENES PROFESIONALES

U. N. A. M.
FACULTAD DE ESTUDIOS SUPERIORES-CUAUTITLÁN

ASUNTO: VOTO APROBATORIO

M. en C. JORGE ALFREDO CUÉLLAR ORDAZ
DIRECTOR DE LA FES CUAUTITLAN
PRESENTE

ATN: I.A. LAURA MARGARITA CORTAZAR FIGUEROA
Jefa del Departamento de Exámenes Profesionales
de la FES Cuautitlán.

Con base en el Reglamento General de Exámenes, y la Dirección de la Facultad, nos permitimos comunicar a usted que revisamos el: **Trabajo de Tesis**

Estudio de sesgos o prejuicios en el aprendizaje computacional: enfoque en el diseño de modelos con clasificadores bayesianos.

Que presenta el pasante: **Mauricio Byrd Victorica**
Con número de cuenta: **312530336** para obtener el título de: **Ingeniero en Telecomunicaciones, Sistemas y Electrónica**

Considerando que dicho trabajo reúne los requisitos necesarios para ser discutido en el **EXAMEN PROFESIONAL** correspondiente, otorgamos nuestro **VOTO APROBATORIO.**

ATENTAMENTE
"POR MI RAZA HABLARÁ EL ESPÍRITU"
Cuautitlán Izcalli, Méx. a 23 de Septiembre de 2020.

PROFESORES QUE INTEGRAN EL JURADO

	NOMBRE	FIRMA
PRESIDENTE	Mtro. Leopoldo Martín del Campo Ramírez	
VOCAL	Ing. José Luis Barbosa Pacheco	
SECRETARIO	Mtro. José Isaac Sánchez Guerra	
1er. SUPLENTE	Lic. Mauricio Jaques Soto	
2do. SUPLENTE	Dr. David Tinoco Varela	

NOTA: los sinodales suplentes están obligados a presentarse el día y hora del Examen Profesional (art. 127).

LMCF/cga*

VNIVERSIDAD NACIONAL AVPNOMA DE MEXICO

FACULTAD DE ESTUDIOS SUPERIORES CUAUTITLÁN
SECRETARÍA GENERAL
DEPARTAMENTO DE EXÁMENES PROFESIONALES

U. N. A. M.
FACULTAD DE ESTUDIOS SUPERIORES-CUAUTITLÁN

ASUNTO: VOTO APROBATORIO

M. en C. JORGE ALFREDO CUÉLLAR ORDAZ
DIRECTOR DE LA FES CUAUTITLAN
PRESENTE

ATN: I.A. LAURA MARGARITA CORTAZAR FIGUEROA
Jefa del Departamento de Exámenes Profesionales
de la FES Cuautitlán.

Con base en el Reglamento General de Exámenes, y la Dirección de la Facultad, nos permitimos comunicar a usted que revisamos el: **Trabajo de Tesis**

Estudio de sesgos o prejuicios en el aprendizaje computacional: enfoque en el diseño de modelos con clasificadores bayesianos.

Que presenta el pasante: **Mauricio Byrd Victorica**
Con número de cuenta: **312530336** para obtener el título de: **Ingeniero en Telecomunicaciones, Sistemas y Electrónica**

Considerando que dicho trabajo reúne los requisitos necesarios para ser discutido en el **EXAMEN PROFESIONAL** correspondiente, otorgamos nuestro **VOTO APROBATORIO.**

ATENTAMENTE
"POR MI RAZA HABLARÁ EL ESPÍRITU"
Cuautitlán Izcalli, Méx. a 23 de Septiembre de 2020.

PROFESORES QUE INTEGRAN EL JURADO

	NOMBRE	FIRMA
PRESIDENTE	Mtro. Leopoldo Martín del Campo Ramírez	
VOCAL	Ing. José Luis Barbosa Pacheco	
SECRETARIO	Mtro. José Isaac Sánchez Guerra	
1er. SUPLENTE	Lic. Mauricio Jaques Soto	
2do. SUPLENTE	Dr. David Tinoco Varela	

NOTA: los sinodales suplentes están obligados a presentarse el día y hora del Examen Profesional (art. 127).

LMCF/cga*

VNIVERSIDAD NACIONAL AVPNOMA DE MEXICO

FACULTAD DE ESTUDIOS SUPERIORES CUAUTITLÁN
SECRETARÍA GENERAL
DEPARTAMENTO DE EXÁMENES PROFESIONALES

U. N. A. M.
FACULTAD DE ESTUDIOS SUPERIORES CUAUTITLÁN

ASUNTO: VOTO APROBATORIO

M. en C. JORGE ALFREDO CUÉLLAR ORDAZ
DIRECTOR DE LA FES CUAUTITLAN
PRESENTE

ATN: I.A. LAURA MARGARITA CORTAZAR FIGUEROA
Jefa del Departamento de Exámenes Profesionales
de la FES Cuautitlán.

Con base en el Reglamento General de Exámenes, y la Dirección de la Facultad, nos permitimos comunicar a usted que revisamos el: **Trabajo de Tesis**

Estudio de sesgos o prejuicios en el aprendizaje computacional: enfoque en el diseño de modelos con clasificadores bayesianos.

Que presenta el pasante: **Mauricio Byrd Victorica**
Con número de cuenta: **312530336** para obtener el título de: **Ingeniero en Telecomunicaciones, Sistemas y Electrónica**

Considerando que dicho trabajo reúne los requisitos necesarios para ser discutido en el **EXAMEN PROFESIONAL** correspondiente, otorgamos nuestro **VOTO APROBATORIO.**

ATENTAMENTE
"POR MI RAZA HABLARÁ EL ESPÍRITU"
Cuautitlán Izcalli, Méx. a 23 de Septiembre de 2020.

PROFESORES QUE INTEGRAN EL JURADO

	NOMBRE	FIRMA
PRESIDENTE	Mtro. Leopoldo Martín del Campo Ramírez	
VOCAL	Ing. José Luis Barbosa Pacheco	
SECRETARIO	Mtro. José Isaac Sánchez Guerra	
1er. SUPLENTE	Lic. Mauricio Jaques Soto	
2do. SUPLENTE	Dr. David Tinoco Varela	

NOTA: los sinodales suplentes están obligados a presentarse el día y hora del Examen Profesional (art. 127).

LMCF/cga*

VNIVERSIDAD NACIONAL AVPNOMA DE MEXICO

FACULTAD DE ESTUDIOS SUPERIORES CUAUTITLÁN
SECRETARÍA GENERAL
DEPARTAMENTO DE EXÁMENES PROFESIONALES

U. N. A. M.
FACULTAD DE ESTUDIOS SUPERIORES-CUAUTITLÁN

ASUNTO: VOTO APROBATORIO

M. en C. JORGE ALFREDO CUÉLLAR ORDAZ
DIRECTOR DE LA FES CUAUTITLAN
PRESENTE

ATN: I.A. LAURA MARGARITA CORTAZAR FIGUEROA
Jefa del Departamento de Exámenes Profesionales
de la FES Cuautitlán.

Con base en el Reglamento General de Exámenes, y la Dirección de la Facultad, nos permitimos comunicar a usted que revisamos el: **Trabajo de Tesis**

Estudio de sesgos o prejuicios en el aprendizaje computacional: enfoque en el diseño de modelos con clasificadores bayesianos.

Que presenta el pasante: **Mauricio Byrd Victorica**
Con número de cuenta: **312530336** para obtener el título de: **Ingeniero en Telecomunicaciones, Sistemas y Electrónica**

Considerando que dicho trabajo reúne los requisitos necesarios para ser discutido en el **EXAMEN PROFESIONAL** correspondiente, otorgamos nuestro **VOTO APROBATORIO.**

ATENTAMENTE
"POR MI RAZA HABLARÁ EL ESPÍRITU"
Cuautitlán Izcalli, Méx. a 23 de Septiembre de 2020.

PROFESORES QUE INTEGRAN EL JURADO

	NOMBRE	FIRMA
PRESIDENTE	Mtro. Leopoldo Martín del Campo Ramírez	
VOCAL	Ing. José Luis Barbosa Pacheco	
SECRETARIO	Mtro. José Isaac Sánchez Guerra	
1er. SUPLENTE	Lic. Mauricio Jaques Soto	
2do. SUPLENTE	Dr. David Tinoco Varela	

NOTA: los sinodales suplentes están obligados a presentarse el día y hora del Examen Profesional (art. 127).

LMCF/cga*

AGRADECIMIENTOS

A mi facultad y universidad, por las muchas oportunidades provistas.

A todos mis auténticos docentes, por su genuina guía, muy por encima de evaluaciones triviales, es por ustedes que esta universidad es tan superior a sus circunstancias. Especialmente agradezco a mi asesor, por lo ya mencionado, el desafío, el compromiso y el apoyo, siempre por encima de las expectativas académicas.

A mis padres, por su apoyo incondicional, el privilegio que recibí por su esfuerzo, y sobre todo, su amor incuestionable.

A Ernesto, por tu valioso compañerismo en los numerosos desafíos académicos (incluyendo este último), pero principalmente por tu invaluable amistad, que lo hizo todo mucho más llevadero. José, Raúl y Alfonso, no podría omitirlos aquí considerando su parte en mi proceso, gracias a todos.

A Andrea, Alejandro, Luisa y Alec. Quizá es inusual agradecerles en este contexto, pero muchos caminos y procesos míos, incluyendo este, serían muy diferentes sin ustedes (para mal).

A Dios, por todos los anteriores, por sostenerme a cada instante, y por ofrecerme más de lo que sé percibir.

ÍNDICE

1. INTRODUCCIÓN

En los últimos años, la toma de decisiones automatizada o asistida mediante técnicas de inteligencia artificial, y particularmente de aprendizaje computacional o automático (*machine learning*), se ha convertido en una parte esencial de numerosas aplicaciones fundamentales en el mundo digital (el cual, por su parte, se ha infiltrado en el corazón del mundo mismo). Estas aplicaciones van desde sistemas de recomendaciones y publicidad personalizada, hasta sistemas predictivos en ámbitos legales, económicos y de salud; las decisiones tomadas van desde lo trivial hasta lo crucial, y conforme el uso de estas técnicas aumenta, crece el interés por garantizar que las mismas sean transparentes, razonables y justas.

El sesgo en el aprendizaje computacional se refiere a la discriminación injustificada por parte de estos sistemas, que pueden incorporar prejuicios humanos en su estructura, amplificarlos y finalmente incurrir en injusticias sistemáticas. En las aplicaciones donde evitar el sesgo es más relevante, los modelos usados suelen ser clasificadores, pues la tarea principal generalmente es clasificar individuos en diversas categorías (personas calificadas y no calificadas para recibir préstamos o empleos, personas diagnosticadas o no con cierta enfermedad, personas que manifiestan o no la probabilidad de cometer crímenes en el futuro, etcétera). A su vez, los modelos clasificadores de aprendizaje computacional pueden ser bayesianos o no, y de serlo, hay varios elementos que deben incorporarse en los modelos, propios de su interpretación probabilística (elementos ausentes en el caso no bayesiano o frecuentista, que suele ser más simple).

El uso de la metodología bayesiana también se ha incrementado desde hace varios años, ya que antes era complicado implementarla debido al costo computacional requerido por las técnicas y procesos de la misma. La interpretación bayesiana de la probabilidad tiene fundamentos teóricos sólidos, ausentes en la interpretación frecuentista u ortodoxa, estrechamente relacionados con el sentido común, la teoría de la toma de decisiones y la lógica misma, que se traducen en valiosas ventajas en las aplicaciones del aprendizaje computacional. En este trabajo se expone la problemática del sesgo o prejuicio, con un enfoque en las implicaciones de involucrar una metodología bayesiana en el diseño de los modelos clasificadores usados para el estudio del sesgo.

El desarrollo del trabajo consiste en analizar un caso de estudio real de una herramienta diseñada para predecir el riesgo de reincidencia delictiva de individuos en el sistema de justicia, y en particular, estudiar la evaluación del sesgo racial presente en la herramienta mediante modelos clasificadores. Se toman dos estudios del mismo problema, donde el segundo es una réplica y objeción al primero. Ambos casos usan una metodología frecuentista para analizar el problema y derivar sus conclusiones. Este trabajo consiste en implementar ambos estudios con el método bayesiano, observando las ventajas y costos adicionales que implica el cambio de método al implementar modelos, obtener resultados y derivar conclusiones. El objetivo no es reconciliar las diferencias entre ambos puntos de vista, o presentar una tercera conclusión diferente para el caso de estudio, sino estudiar lo que la metodología bayesiana puede ofrecer en el desarrollo independiente de cada uno. Los capítulos del trabajo se pueden agrupar en dos partes: la fundamentación teórica (capítulos 2-5) y el desarrollo (capítulos 6-12).

La primera parte comienza en el capítulo 2, con notaciones y conceptos matemáticos fundamentales, persistentes en el escrito. En esta primera parte también se desarrollan conceptos y ejemplos pertinentes para definir: el método bayesiano (y sus diferencias con el método frecuentista) en el capítulo 3, la discriminación y su análisis (en el contexto de la estadística y el aprendizaje computacional) en el capítulo 4, y los modelos clasificadores del aprendizaje computacional en el capítulo 5.

En la segunda parte se comienza con la descripción y justificación de las herramientas usadas para el desarrollo (capítulo 6), y después se desarrolla un caso con datos sintéticos para ejemplificar y describir varios conceptos en un sentido más práctico (capítulo 7). Posteriormente se hace una breve descripción del estudio del sesgo en el aprendizaje computacional, planteando también el estado del arte (capítulo 8), seguida del planteamiento del caso de estudio a analizar y el

trabajo relacionado a este escrito (capítulo 9). Finalmente se pasa a las partes principales del desarrollo, donde se presentan las implementaciones de ambos casos de estudio, incluyendo: reproducciones de los trabajos originales, diseños e implementaciones para los modelos originales y los modelos bayesianos, y los resultados obtenidos con su respectivo análisis. El primer estudio se desarrolla en el capítulo 10, y el segundo en el capítulo 11. Finalmente se cierra con las conclusiones, incluyendo sugerencias para trabajo a futuro en el capítulo 12. El anexo del trabajo contiene varios resultados que no están dentro del enfoque principal del escrito, pero son complementos interesantes e importantes para el desarrollo central de los análisis del caso de estudio.

En lugar de dedicar una sección para todas las referencias del escrito, en cada capítulo las referencias pertinentes se incluyen al final del mismo, con la intención de facilitar al lector la consulta de las mismas. Todo el código desarrollado para este escrito está documentado y disponible en el siguiente enlace: https://github.com/gerkbyrd/tesis_sesgo_ML_clasificadores_bayesianos.

1.1 PREGUNTA DE INVESTIGACIÓN Y OBJETIVOS

Pregunta de Investigación

¿Cuáles son las implicaciones de usar clasificadores bayesianos en el diseño de modelos aplicados al estudio del sesgo en el aprendizaje computacional, especialmente en contraste con sus contrapartes frecuentistas?

Objetivo General

Demostrar y analizar el impacto de usar la metodología bayesiana en el diseño de modelos clasificadores de aprendizaje computacional, aplicados al estudio del sesgo en la toma automática de decisiones.

Objetivos Específicos

- Presentar una comparación general entre las metodologías bayesiana y frecuentista en el contexto de clasificadores de aprendizaje computacional, y en particular, en el modelo de regresión logística.
- Desarrollar la comparación enfocada al estudio del sesgo en un caso de estudio real, presentando el diseño e implementación de los modelos usados, mediante ambas metodologías, para comparar y analizar los resultados.
- Desarrollar la comparación entre metodologías sobre un segundo caso real, estudiando el mismo problema del primer caso, pero con un planteamiento que lleva a conclusiones contrarias, exponiendo nuevamente el diseño e implementación de los modelos necesarios, para analizar las diferencias entre metodologías.
- De los resultados obtenidos, derivar conclusiones generales sobre el uso de la metodología bayesiana en el estudio del sesgo, puntualizando las limitaciones del presente trabajo y sugiriendo nuevas direcciones posibles para el mismo.

2. NOTACIÓN Y CONCEPTOS IMPORTANTES

Aprendizaje Computacional: este es el término en español empleado para referirse a *machine learning*, que es posiblemente el área de estudio más popular en la actualidad dentro de la inteligencia artificial. Murphy (2012, p. 1) lo define como un conjunto de métodos capaz de detectar automáticamente patrones en los datos, y mediante dichos patrones, predecir futuros datos o realizar otros tipos de toma de decisiones bajo incertidumbre. Actualmente otra traducción popular es "aprendizaje automático", pero se optó por "aprendizaje computacional", por ser el término usado por el laboratorio de MLPR (*Machine Learning and Pattern Recognition*) del INAOE, o en español: laboratorio de aprendizaje computacional y reconocimiento de patrones (*INAOE - Ciencias Computacionales*, s. f.).

Sesgo: el sesgo, dentro del aprendizaje computacional, puede referirse al prejuicio contra grupos o individuos en aplicaciones («Our Mission», 2016), que es el caso en el título de este trabajo. Sin embargo, hay otro concepto básico dentro del aprendizaje computacional con el mismo nombre, también presente en este trabajo (se aclara la distinción cuando es necesario): el sesgo o *bias* se refiere al término constante en la regresión de las funciones que se aproximan o se aprenden. En ese contexto, no tiene nada que ver con prejuicios o injusticias, es meramente un elemento matemático (Murphy, 2012, p. 20).

Clasificadores Bayesianos: otra aclaración necesaria es la del término de "clasificadores bayesianos". El uso del término, en el título del trabajo, se refiere a cualquier tipo de clasificador usado con una metodología bayesiana. El otro uso se refiere al tipo de clasificadores que emplean el teorema de Bayes dentro de su estructura y algoritmo. Por ahora es suficiente decir que una regresión logística bayesiana (Murphy, 2012, p. 254), es un clasificador bayesiano en la forma a que se refiere el título del trabajo, mientras que un clasificador bayesiano (o de Bayes) ingenuo (*Naive Bayes*) (Murphy, 2012, pp. 82-83), es bayesiano en el segundo significado de la palabra. Por esta diferencia, puede hablarse de “clasificadores bayesianos de Bayes”, pues es posible tener clasificadores que son bayesianos en ambos sentidos de la palabra. Este último caso no se discute más en el resto de este trabajo (Murphy, 2012, pp. 84-85).

Aunque la notación empleada no es tan extensiva como para merecer anexos similares a los que hay en los libros de texto sobre el tema, aclarar algunos elementos básicos agilizará la narrativa en los siguientes capítulos, evitando la necesidad de interrumpir pasajes para explicar la notación.

Álgebra Lineal:

Vectores: constantemente se hará uso de vectores para expresar entradas, salidas, variables, etc. Estos se escribirán en negritas, y al expandirlos en sus componentes estarán entre corchetes. Se suele tratar con vectores "columna", que tienen una forma vertical. Una "T" en el exponente quiere decir que el vector (o matriz) está transpuesto. Por ejemplo, el vector columna $\mathbf{x}$ con componentes x_1, x_2, x_3 se expresa como:

$$\mathbf{x} = [x_1, x_2, x_3]^{\mathrm{T}}$$

Equivalente a:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

A su vez, los escalares se expresan en cursivas, como los componentes x_i del vector $\mathbf{x}$ aquí.

Matrices: por su parte, las matrices se expresarán con letras mayúsculas. Por ejemplo:

$$\mathrm{X} = \begin{bmatrix} x_{11} x_{12} x_{13} \\ x_{21} x_{22} x_{23} \\ x_{31} x_{32} x_{33} \end{bmatrix}$$

Productos: en este trabajo los productos entre cualquier combinación de matrices y vectores serán siempre escalares o productos punto; para simplificar la notación se omite el punto y simplemente se escriben los elementos de forma contigua, por ejemplo:

Considerando:

$$\mathbf{x} = [x_1, x_2, x_3]^{\mathrm{T}}$$
$$\mathbf{w} = [w_1, w_2, w_3]^{\mathrm{T}}$$

Se tiene:

$$\mathbf{w}^{\mathbf{T}}\mathbf{x} = \begin{bmatrix} w_1 w_2 w_3 \end{bmatrix} . \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = w_1 x_1 + w_2 x_2 + w_3 x_3$$

Cualquier discusión con respecto a los conceptos de álgebra lineal usados en este trabajo está fuera del enfoque del mismo. Para las definiciones implícitas en la notación presentada, y el trabajo en sí, se puede tomar como referencia (Barber, s. f.), donde estos conceptos se presentan con mayor detalle.

Probabilidad y Estadística:

Las probabilidades se definen como cantidades numéricas no negativas, definidas sobre una serie de resultados posibles, aditivas sobre resultados mutuamente exclusivos, y que suman a 1 sobre todos los resultados mutuamente exclusivos posibles (Murphy, 2012, pp. 27-28).

En cuanto a su notación, se tendrá "P" para probabilidades discretas y "*p*" para distribuciones de probabilidad continuas, también llamadas densidades de probabilidad o "PDF" (por las siglas en inglés para "función de densidad de probabilidad"). A su vez, se emplearán comas para probabilidades conjuntas y se usará " | " para probabilidades condicionales, es decir:

Probabilidad de A = P(A); Probabilidad de B = P(B); Probabilidad de C = P(C).

Probabilidad de A y B = P(A, B); Probabilidad de A, B y C = P(A, B, C).

Probabilidad de B dado A: P(B | A); Probabilidad de A, dado B y C = P(A | B, C) .

Los ejemplos del primer renglón, son de **probabilidades marginales**, los del segundo son de **probabilidades conjuntas** y los del tercer renglón son de **probabilidades condicionales** (Murphy, 2012, p. 29). En estos ejemplos se usan probabilidades discretas, pero la notación persiste para casos continuos donde se usan PDFs en su lugar.

Los siguientes conceptos son relevantes para el tema, y aunque un tratamiento explícito de los mismos no es crucial para el desarrollo del escrito, es importante tenerlos presentes. La referencia para estos conceptos es (Gelman et al., 2013, pp. 4-6).

Inferencia: es el objetivo central de la estadística, que consiste en sacar conclusiones acerca de datos no observados a partir de datos observados. La inferencia causal es la comparación entre el resultado observado bajo ciertas condiciones y el resultado que no se observó porque se hubiera dado bajo condiciones diferentes.

Cantidad de interés ("estimando"): son las cantidades no observadas, los objetos de la inferencia. Pueden ser potencialmente observables (como datos futuros) o no directamente observables (parámetros del proceso hipotético que explica los datos).

Intercambiabilidad: es una suposición básica de todo análisis estadístico, y significa que los valores observados del estimando (o cualquier variable aleatoria) son igual de probables independientemente del ordenado de los mismos. Los datos de distribuciones de probabilidad con esta propiedad, suelen modelarse como independientes e idénticamente distribuidos (iid).

Variables explicativas ("predictores"): en cada unidad o registro presente en los datos, hay observaciones que no se modelan como aleatorias. Una unidad consiste de valores de entrada y salida. Las variables explicativas (o covariables) son aquellas que se consideran entradas en las unidades. La intercambiabilidad puede extenderse a estas variables, de modo que dos unidades con las mismas entradas tengan la misma salida o distribución para el estimando.

Modelos jerárquicos: se usan cuando hay información disponible en diferentes niveles de unidades observacionales, considerando intercambiabilidad dentro de cada nivel[1].

REFERENCIAS

Barber, D. (s. f.). Learning from Data Supplementary Mathematics (Vector and Linear Algebra). *Learning from Data*, 32.

Gelman, A., Stern, H. S., Carlin, J. B., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. Chapman and Hall/CRC.

INAOE - Ciencias Computacionales. (s. f.). Recuperado 22 de mayo de 2020, de https://ccc.inaoep.mx/laboratorios/mlpr.php

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT press.

Our Mission. (2016, julio 8). *UnBias*. https://unbias.wp.horizon.ac.uk/our-mission/

1 Un ejemplo que ilustra con claridad los modelos jerárquicos se presenta en (Gelman et al., 2013, pp. 102-104), donde el modelo de un experimento para estimar la probabilidad de tumores en ratas que no reciben un tratamiento, se convierte en jerárquico al haber múltiples experimentos, cada uno con su propio grupo de ratas. En este caso las ratas en cada experimento son unidades intercambiables en el primer nivel, mientras que los experimentos son unidades intercambiables en el segundo nivel.

3. EL MÉTODO BAYESIANO: DEFINICIÓN Y JUSTIFICACIÓN

El objetivo de este capítulo es definir y justificar la metodología bayesiana, pues la presente tesis se enfoca en los clasificadores bayesianos como sistemas de toma de decisiones. Un buen punto de partida es puntualizar las diferencias entre el enfoque bayesiano y la alternativa que fue mayoritariamente preferida durante décadas: la estadística frecuentista u ortodoxa. Partiendo de este contraste, se puede plantear la justificación para optar por el método bayesiano, y comprender entonces su potencial dentro de la aplicación del sesgo en el aprendizaje computacional. En las demás secciones se expande la definición de la metodología, tratando los conceptos esenciales para este trabajo.

3.1 EL DEBATE ENTRE LAS ESTADÍSTICAS FRECUENTISTA Y BAYESIANA

La **estadística bayesiana** se basa en emplear una distribución probabilística a posteriori $p(\boldsymbol{\theta} \mid \mathbf{D})$ para resumir lo que se sabe de un conjunto de variables desconocidas. Esta distribución se puede representar con un "resumen estadístico"[2] derivado de la misma, más fácil de manipular y visualizar (Murphy, 2012, p. 149). La oposición al uso del teorema de Bayes en problemas estadísticos se debe a que considera que los parámetros son variables aleatorias, tomando distintos valores con cierta probabilidad para cada uno. La **estadística frecuentista** considera que los parámetros tienen un valor fijo y desconocido. En vez de una distribución a posteriori, se lidia con una distribución muestreada para un "estimador", obtenida al aplicar dicho estimador sobre varios conjuntos de datos (*datasets*), muestreados de la verdadera distribución desconocida. En ese caso la incertidumbre se modela con base en la variación entre pruebas o experimentos repetidos (Murphy, 2012, p. 191).

La metodología frecuentista se basa en la **reducción de riesgo empírica**, que emplea los datos disponibles para estimar la distribución desconocida, y así minimizar alguna función de pérdida que denote la diferencia entre el valor verdadero y el valor estimado. En el método bayesiano, se condiciona sobre los datos observados (no existe la noción de pruebas o experimentos repetidos), permitiendo calcular la probabilidad de eventos únicos, y evitando paradojas propias del método frecuentista (que se revisarán más adelante). A pesar de haber fuertes argumentos teóricos a favor de la estadística bayesiana, no se prefiere de forma unánime, y el método frecuentista es ampliamente usado dentro del aprendizaje computacional (Murphy, 2012, pp. 191, 204-205, 215).

Otra forma de describir la diferencia es la siguiente: en la **estadística frecuentista**, se construyen "estimadores" de las cantidades de interés, y se elige entre ellos mediante algún criterio de sus propiedades de muestreo; no hay principios generales para elegir el criterio, y por lo general tampoco hay un procedimiento sistemático para construir un estimador óptimo. En la **estadística bayesiana**, una vez que se hacen explícitas las suposiciones del modelo y los datos, la inferencia es un proceso "mecánico". La teoría de la probabilidad dará respuestas únicas considerando toda la información disponible. No hay necesidad de *diseñar* estimadores ni intervalos de confianza, sólo se debe diseñar el espacio de hipótesis (las especificaciones del modelo y sus distribuciones de probabilidad) y determinar cómo implementar la inferencia en dicho espacio (en el sentido computacional). Como resultado se tienen distribuciones sobre las cantidades de interés: la estadística frecuentista devuelve como resultado su "mejor conjetura" para la variable dependiente en alguna ubicación del espacio de la variable independiente, mientras que la bayesiana da una distribución de los valores posibles de la variable dependiente en cada ubicación del espacio de la variable independiente. Es decir, en el caso bayesiano no hay necesidad de comparar entre tantos valores distintos de los parámetros del modelo, y se puede especificar la incertidumbre que se tiene con respecto al modelo y sus parámetros; se usan distribuciones para representar creencias, y la teoría de la probabilidad para actualizar las mismas. La incertidumbre no es más que la varianza de dichas distribuciones (*MLPR w6b - Machine Learning and Pattern Recognition*, s. f.), (MacKay, 2003, p. 320).

2 Es un resumen de las observaciones para describirlas de la forma más simple posible, algunos resúmenes estadísticos son la media, la moda, y la varianza («Statistics», s. f.).

El hecho de que uno de los mayores obstáculos para la estadística bayesiana sea la carga computacional, hace que el enfoque se vuelva más atractivo con el avance tecnológico continuo. El método frecuentista es atractivo en casos donde la actualización bayesiana es costosa o ineficaz, o donde las suposiciones con respecto al problema son débiles o incluso adversativas[3] (Steinhardt, 2012).

3.2 METODOLOGÍA BAYESIANA

Más allá de las ventajas o desventajas que tenga una metodología bayesiana ante una frecuentista, es importante definir con mayor detalle en qué consiste. Es un tema amplio, pero con base en algunos trabajos fundamentales ((Gelman et al., 2013), (MacKay, 2003), (Murphy, 2012)) se establecerán los fundamentos necesarios para discutir los modelos planteados en este trabajo.

Murphy (2012) ilustra con claridad varios puntos importantes de la metodología bayesiana, basándose en el contraste con la estadística frecuentista, y hablando de los argumentos teóricos a favor del método bayesiano, cuya ausencia en el método ortodoxo son "fallas fundamentales" y "paradojas" ligadas al mismo. Entre estas fallas sobresale la violación del principio de la verosimilitud, que establece que las inferencias deberían basarse en la verosimilitud de los datos observados (como en el método bayesiano) y no en datos hipotéticos del futuro (como en el método ortodoxo). Este principio, a su vez, se basa en el principio de la suficiencia, el cual establece que una "estadística suficiente" contiene toda la información relevante de algún parámetro desconocido. Y además, se basa en el principio de la condicionalidad débil, que dicta que las inferencias deben basarse en eventos que pasaron, no que pudieran haber pasado (Murphy, 2012, pp. 214-215).

En cuanto al método en sí, en (Gelman et al., 2013) el análisis bayesiano de datos se resume en tres pasos (Gelman et al., 2013, p. 3):

1. Establecer un modelo de probabilidad completo: una distribución conjunta para todas las cantidades (observables o no) del problema. Éste debe ser consistente con el conocimiento acerca del problema y el proceso de recolección de datos.
2. Condicionar sobre datos observados: calcular e interpretar la distribución a posteriori adecuada, que es la distribución condicional de las cantidades de interés (que no se han observado), dados los datos observados.
3. Evaluar el rendimiento del modelo y las implicaciones de la distribución a posteriori: no se limita a cuán preciso es el modelo con respecto a los datos, también involucra evaluar la factibilidad de las conclusiones y la sensatez de los resultados, considerando las suposiciones del modelado. En este paso el modelo puede se alterado o expandido, para repetir el proceso desde el primer paso.

Se ha mencionado una relación entre la escuela bayesiana y el sentido común, que se ve reflejada en la interpretación de las conclusiones estadísticas. Un intervalo de probabilidad bayesiana es aquél que tiene una alta probabilidad de contener el valor de interés[4]. En un intervalo de confianza frecuentista, el significado está sujeto a secuencias de inferencias similares y repetibles. En la estadística aplicada, cada vez es más común la estimación de intervalos que las pruebas de hipótesis, y encima de esto, los intervalos de confianza frecuentistas suelen ser interpretados

3 Una suposición adversativa se refiere a asumir un contexto donde un adversario tiene acceso total al modelo y busca oponerse al objetivo del mismo. Por ejemplo, si el modelo se usa para apostar contra algún oponente, la situación adversativa es aquella donde dicho oponente tiene acceso total al modelo. En (Steinhardt, 2012) se menciona que las suposiciones explícitas del método bayesiano son desfavorables en este contexto, pues aunque generalmente es bueno que quien conozca las suposiciones llegue a los mismos resultados, si se trata de un adversario, se convierte en una severa desventaja, y por eso es una de las posibles ventajas del método frecuentista.

4 Regularmente se buscan intervalos con probabilidad alta porque se puede decir con certeza que el valor en cuestión está ahí, pero técnicamente se pueden establecer intervalos con cualquier probabilidad de contener el valor.

erróneamente con el significado de los intervalos de probabilidad bayesiana (Gelman et al., 2013, p. 3).

También se ha hablado de la capacidad bayesiana de "cuantificar la incertidumbre". La utilidad no se limita a representar numéricamente la certeza de los resultados o predicciones; esta cualidad implica que, en teoría, no hay ningún impedimento para emplear modelos con muchos parámetros y especificaciones probabilísticas complicadas o de varias capas. La metodología bayesiana permite abordar problemas complejos con importantes ventajas pragmáticas. La simpleza conceptual bayesiana facilita la libertad de construir modelos complejos (Gelman et al., 2013, p. 4).

Inferencia Bayesiana: las conclusiones bayesianas se expresan en términos de declaraciones probabilísticas, condicionadas sobre el valor observado del "estimando" (valor de interés) y sobre los valores conocidos de las covariables (valores de entrada). Este condicionamiento sobre los datos observados distingue a la estadística bayesiana del caso ortodoxo, donde la base es la evaluación retrospectiva de la estimación de los parámetros, sobre la distribución de los posibles valores de salida, donde dicha distribución está condicionada sobre el valor verdadero de los parámetros (Gelman et al., 2013, p. 6).

La intercambiabilidad en el contexto bayesiano quiere decir que la incertidumbre se expresa como una distribución de probabilidad conjunta, invariante con el orden de las probabilidades marginales del conjunto (Gelman et al., 2013, p. 5).

Los siguientes conceptos son definiciones generales y se pueden consultar de cualquier fuente en el tema, pero a continuación se presentan como se describen en (Gelman et al., 2013, pp. 6-8):

Teorema de Bayes: considerando los parámetros del modelo $\boldsymbol{\theta}$, y un estimando y, el teorema expresa la probabilidad de $\boldsymbol{\theta}$ dado y. Primero es importante tener en cuenta la regla del producto, que expresa la probabilidad conjunta de dos variables:

$$p(\boldsymbol{\theta}, y) = p(\boldsymbol{\theta})p(y \mid \boldsymbol{\theta}) = p(y)p(\boldsymbol{\theta} \mid y)$$

El Teorema de Bayes es un simple despeje de la segunda igualdad:

$$p(\boldsymbol{\theta}|y) = \frac{p(\boldsymbol{\theta})\, p(y|\boldsymbol{\theta})}{p(y)}$$

Esta es la **probabilidad a posteriori**. Es importante la forma no normalizada, obtenida al omitir el único término que no depende de los parámetros $\boldsymbol{\theta}$:

$$p(\boldsymbol{\theta}|y) \propto p(\boldsymbol{\theta})\, p(y|\boldsymbol{\theta})$$

El término $p(y \mid \boldsymbol{\theta})$, a pesar de ser la probabilidad de y dado $\boldsymbol{\theta}$, se considera como una función de $\boldsymbol{\theta}$, denominada la **verosimilitud** de $\boldsymbol{\theta}$. Esta función es crucial, ya que es la forma en que los datos observados afectan la probabilidad a posteriori. Como el objetivo es inferir los parámetros a partir de los datos observados, este término se interpreta como "verosimilitud de $\boldsymbol{\theta}$" y no "probabilidad de y".

Como ya se mencionó antes, la inferencia bayesiana obedece el **principio de la verosimilitud**: para una muestra de datos, cualquiera de los modelos de probabilidad $p(y \mid \boldsymbol{\theta})$ con la

misma función de verosimilitud, dará la misma inferencia de los parámetros $\boldsymbol{\theta}$. El principio es razonable dentro de la familia de modelos que se adopte para un análisis particular.

El término $p(\boldsymbol{\theta})$ es la **probabilidad a priori**. Su nombre se debe a que no está condicionada sobre observaciones previas. Lo ideal es que esta distribución represente los conocimientos previos del proceso, los cuales están presentes antes de hacer inferencias con los datos.

Midiendo la Incertidumbre

En el método bayesiano, la probabilidad es la medida fundamental de la incertidumbre. Al separarse de los requerimientos de reproducibilidad y repetitividad del método frecuentista[5], se puede lidiar con otro tipo de situaciones no necesariamente repetibles (la probabilidad de que una película gane un premio, o que un atleta gane algún torneo, por ejemplo). Los métodos bayesianos permiten hacer declaraciones de forma sistemática acerca del conocimiento parcial disponible con respecto a alguna situación, mediante el uso de la probabilidad; el estado actual del conocimiento de algo desconocido es descrito con una distribución de probabilidad (Gelman et al., 2013, pp. 11-12).

En (Gelman et al., 2013) se citan algunos argumentos comunes a favor de usar la probabilidad para cuantificar la incertidumbre: la aleatoriedad física induce incertidumbre, por lo que es razonable describir la incertidumbre en el lenguaje de eventos aleatorios (que es el lenguaje de la probabilidad). Con base en la teoría de la decisión, considerando toda inferencia estadística dentro de ese contexto, los axiomas correspondientes a esta teoría implican que la incertidumbre debe representarse en términos probabilísticos. Con el principio de coherencia aplicado a las apuestas, donde se asignan probabilidades con el fin de no permitir ganancias definitivas al rival, la construcción de dichas probabilidades deben satisfacer los axiomas de la teoría de la probabilidad. A pesar de estos argumentos, el autor concluye que la prueba definitiva debe depender del éxito en las aplicaciones (Gelman et al., 2013, pp. 12-13).

Hablando en un sentido más práctico, y que se verá reflejado en los experimentos realizados más adelante en este escrito, al representar creencias con distribuciones probabilísticas, la varianza es la cantidad que cuantificará directamente la incertidumbre (*MLPR w6b - Machine Learning and Pattern Recognition*, s. f.).

Estadística Bayesiana Aplicada

En el contexto de las aplicaciones, las grandes fortalezas del método bayesiano son: la capacidad de combinar múltiples fuentes de información (lo cual ofrece la posibilidad de conclusiones finales más objetivas[6]), y la consideración de la incertidumbre con respecto a las incógnitas del problema. También es importante la manera en que los modelos bayesianos pueden adaptarse a estructuras

5 En el pensamiento frecuentista la probabilidad está definida como una frecuencia para un evento repetido. La interpretación bayesiana, donde se ve a la probabilidad como un valor esperado razonable, es lo que permite esta cuantificación de la incertidumbre mediante la probabilidad, asociada siempre a la estadística bayesiana. Los fundamentos teóricos involucran áreas como teoría de la decisión, teoría de juegos e incluso lógica booleana. Cox (1946) hace un tratamiento del tema más detallado, pero relativamente conciso, mientras que Savage (1954) provee un trabajo profundo que involucra pruebas matemáticas en un escrito más riguroso, que también es una de las referencias más importantes con respecto a los fundamentos de la estadística bayesiana.

6 Usar idealizaciones matemáticas del mundo siempre implica subjetividad (Gelman et al., 2013, p. 13), pero el método bayesiano, aunque depende de una distribución a priori subjetiva (la probabilidad bayesiana es llamada incluso probabilidad subjetiva o personal (Savage, 1954)), no sólo permite incorporar varias fuentes de información, también es explícito con respecto a su "subjetividad", reflejando así las variaciones que diferentes suposiciones razonables pudieran tener en sus respuestas, y mostrando cuán adecuadas son las suposiciones mismas a la luz de la evidencia observada.

complejas, manteniendo la cohesión en los resultados, así como la manera intuitiva en que pueden interpretarse dichos resultados (Gelman et al., 2013, p. 24).

Otros elementos importantes en la estadística aplicada son: tener disposición a usar muchos parámetros, estructurar los modelos de forma jerárquica, evaluar modelos (en rendimiento y en concepto), hacer inferencias en forma de distribuciones de probabilidad (no sólo estimaciones puntuales), incluir toda la información previa posible, diseñar para que las inferencias sean "robustas" ante las suposiciones del modelo, etc. Estos elementos no son exclusivos para métodos formalmente bayesianos, pero las posibilidades bayesianas de lidiar con estas tareas son argumentos a favor del método desde la perspectiva de la estadística aplicada (Gelman et al., 2013, p. 25).

REFERENCIAS

Cox, R. T. (1946). Probability, Frequency and Reasonable Expectation. *American Journal of Physics*, *14*(1), 1-13. https://doi.org/10.1119/1.1990764

Gelman, A., Stern, H. S., Carlin, J. B., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. Chapman and Hall/CRC.

MacKay, D. J. C. (2003). *Information Theory, Inference, and Learning Algorithms*. Copyright Cambridge University Press.

MLPR w6b—Machine Learning and Pattern Recognition. (s. f.). Recuperado 9 de octubre de 2019, de https://www.inf.ed.ac.uk/teaching/courses/mlpr/2018/notes/w6b_bayesian_regression.html

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT press.

Savage, L. (1954). *The Foundations of Statistics*. Wiley.

Statistics: A Brief Guide | Summarising Data. (s. f.). *Data Analytics*. Recuperado 21 de mayo de 2020, de https://www.dataanalytics.org.uk/data-analytics-knowledge-base-tips-tricks-r-excel/statistics-guide/data-summary/

Steinhardt, J. (2012). *Beyond Bayesians and Frequentists*.

4. ANÁLISIS DE LA DISCRIMINACIÓN

El presente trabajo gira en torno a una metodología (la estadística bayesiana) y una aplicación (el sesgo o prejuicio dentro del aprendizaje computacional). Con respecto a la metodología, sus fundamentos se cubrieron en el capítulo pasado, en el presente se hará lo propio para la aplicación. Normalmente una descripción del problema a tratar, justificando la importancia del mismo, es suficiente (y se hará al entrar en detalle sobre el caso de estudio específico, en el cual se basan las implementaciones y experimentos realizados); sin embargo, al hablar de términos como discriminación o justicia, una descripción del problema puede desviarse fácilmente hacia áreas como el derecho o la filosofía, que a pesar de involucrar discusiones fascinantes, quedan fuera del enfoque del presente trabajo.

Romei & Ruggieri (2014) se dieron a la tarea de elaborar una recopilación multidisciplinaria sobre el análisis de la discriminación[7], donde reúnen los conceptos esenciales en la materia, con un enfoque en los métodos de colección y análisis de datos[8]. Usando su trabajo como referencia, la idea es no desviar la discusión para establecer definiciones de justicia o discriminación, y en cambio usar dicho trabajo como el fundamento con respecto a este tipo de conceptos; que se exponen a continuación acorde a esta referencia (con algunos complementos pertinentes de otras fuentes).

En cuánto a las citas y referencias específicas de esta sección, y con el fin de no saturar el texto innecesariamente, debe asumirse que los conceptos presentados provienen específicamente de (Romei & Ruggieri, 2014, pp. 2–6). Se añadirán referencias sólo para ubicar citas textuales o elementos de otros materiales, hasta la siguiente sección, que trata con las herramientas estadísticas para el análisis de la discriminación.

Discriminación: diferencia injustificada de tratos con base en la pertenencia (percibida o real) a **grupos protegidos** (definidos por rasgos físicos o culturales). Un grupo se considera **protegido** acorde a leyes de derechos humanos si es discriminado tradicionalmente, pero el término puede usarse para cualquier grupo discriminado dependiendo del contexto. Hay excepciones legítimas a la regla para efectos legales (por ejemplo, la discriminación por edad para poder votar en un proceso electoral).

Concretamente, las leyes de los derechos humanos prohíben la discriminación basada en sexo, género, orientación sexual, raza, grupo étnico, color de piel, origen social, rasgos genéticos, lenguaje, religión, creencia, opiniones políticas y personales, pertenencia a minorías nacionales, propiedad, nacimiento, ascendencia, discapacidad, enfermedad, estado civil o edad.

Discriminación múltiple: si alguien es miembro de más de un grupo protegido, puede ser discriminado con base en su pertenencia a cualquier grupo o a la combinación de éstos (discriminación por distintas características en diferentes eventos). Cada discriminación es diferente, y la discriminación múltiple es el impacto acumulativo. Si las distintas discriminaciones sobre alguien tienen un efecto aditivo (ambas ocurren al mismo tiempo, en contraste con el caso múltiple), se denomina **discriminación compuesta**. A su vez, las discriminaciones concurrentes que dan lugar a una forma específica y diferente de discriminación, distinta a cualquiera de los grupos concurrentes en aislamiento, entran en la categoría de **discriminación interseccional**.

7 Específicamente, ellos indican que este problema involucra causas sociológicas, argumentaciones legales, técnicas estadísticas, y problemas computacionales (Romei & Ruggieri, 2014, p. 1).

8 Los autores mencionan que en los principios del diseño estadístico hay métodos observacionales, experimentales y cuasi-experimentales, dependiendo de cómo se coleccionan los datos (observando datos existentes o históricos, u obteniendo nuevos datos con un control parcial o total de los factores): los más presentes en el análisis de la discriminación son los observacionales. Los autores consideran un cuarto método, que será el de mayor interés para este trabajo: el descubrimiento de conocimiento (*knowledge discovery*) (Romei & Ruggieri, 2014, pp. 1, 6, 38).

Es esencial poder evaluar la presencia, naturaleza, extensión y tendencias de la discriminación, así como poder prevenirla en la toma de decisiones (incluyendo la toma automatizada de decisiones, donde está el enfoque de esta tesis) (Romei & Ruggieri, 2014, p. 1). Antes de pasar a los tipos de discriminación, basados en la manera en que se da la misma, la siguiente afirmación es una introducción adecuada:

> *"La discriminación se da únicamente al aplicar reglas y/o prácticas diferentes a situaciones comparables, o reglas y/o prácticas iguales a situaciones diferentes"* (Tobler, 2008, p. 20).

Discriminación directa o sistemática (trato desigual): se da cuando las reglas y/o prácticas afectan explícitamente a alguien desfavorablemente sin una justificación legítima. Es intencional y dirigida (por lo general con base en rasgos físicos). Es difícil de probar, pues las víctimas deben demostrar la intención discriminatoria.

Discriminación indirecta (impacto desigual): es una provisión, práctica o criterio, aparentemente neutral, que resulta en el trato injusto hacia un grupo protegido. Estos elementos "neutrales" tomarán en cuenta atributos personales **correlacionados** con indicadores de características protegidas (sexo raza, etc.). Las leyes correspondientes pretenden evitar la evasión de la prohibición a la discriminación (aún en casos no intencionales).

Se considera presente tras observar los efectos desfavorables de reglas y/o prácticas sobre grupos protegidos, y corresponde al infractor demostrar la ausencia de discriminación, dando una justificación razonable y objetiva.

Justificaciones objetivas: legalmente se proveen en leyes anti-discriminatorias. En la discriminación directa existe el **requerimiento ocupacional genuino**, que significa que la capacidad para algún trabajo se relaciona directamente con características protegidas[9]. Para la discriminación indirecta, basta con justificar que el proceso que provocó el impacto desigual tiene fines legítimos, logrados o trabajados con medios apropiados y necesarios (Ellis, 2005).

Tokenismo y tokenismo inverso: son justificaciones no objetivas para evitar acusaciones de discriminación. El *tokenismo (Kanter, 1977)* consiste en beneficiar a algunos miembros del grupo protegido (*tokens*) para crear una ilusión de inclusión. El caso inverso consiste en negar beneficios injustamente a miembros del grupo mayoritario (*tokens*) para aparentar igualdad.

Causas de la discriminación (en función de la información acerca de los grupos protegidos):

Prejuicio: no hay información acerca de los individuos ni su grupo. El prejuicio resulta en discriminación cuando conlleva actitudes negativas, formadas de manera injusta o irracional, hacia grupos protegidos. El prejuicio fácilmente cae en un círculo vicioso[10], donde causa desventajas sociales que se convierten en evidencia de la inferioridad de dicho grupo, renovando y fortaleciendo el prejuicio original.

Pensamiento estadístico (racismo racional): no hay información acerca de los individuos, pero sí del rendimiento promedio del grupo. La falta de conocimiento acerca de cada individuo del grupo será compensado con el conocimiento del rendimiento del grupo. También se da cuando se toma en cuenta el posible prejuicio contra el grupo del individuo, pero como una característica indeseable del individuo, es decir, quien toma las decisiones no

9 Características que definen a algún grupo protegido (raza, género, etc.).

10 En (Romei & Ruggieri, 2014) los autores citan directamente a Newman (2014) al describir el concepto.

discrimina al individuo porque tenga un prejuicio contra él, sino porque quiere evitar las consecuencias negativas del prejuicio que otros podrían tener (Harford, 2008).

Ignorancia (discriminación no intencional): hay información acerca de los individuos, pero no de los efectos de las decisiones tomadas. No hay intenciones maliciosas, pero no hay suficiente consciencia con respecto a las consecuencias de las acciones (indiferencia, incorrecta ejecución de procedimientos, y falta de planeación y de análisis de consecuencias). La discriminación indirecta y no intencional son problemáticas notables para tomar en cuenta los efectos de las decisiones, pues los efectos discriminatorios son impredecibles.

Acciones afirmativas (o positivas): políticas para compensar, mediante la provisión de oportunidades, a los grupos tradicionalmente afectados. Van desde simplemente alentar, hasta implementar tratamientos y/o cuotas preferenciales. Es importante evaluar y monitorear su aplicación para evitar discriminación inversa (Sowell, 2004).

Discriminación inversa: a veces se asocia con acciones afirmativas, como la desventaja que sufren los miembros de grupos no protegidos a causa de dichas acciones. Para evitarla, se establece que dichas acciones afirmativas no deben mantener derechos desiguales o separados entre grupos raciales, una vez que los objetivos de las mismas se hayan cumplido (The United Nations, 1966).

Favoritismo (nepotismo): se da cuando hay trato preferencial por razones no relacionadas con mérito individual, necesidad de negocios, ni acciones afirmativas. Es la contraparte de la discriminación: un grupo favorecido implica que otro fue discriminado y viceversa. Sin embargo, el término se usa para grupos favorecidos explícitamente, más que para grupos exentos de la discriminación que sufren otros.

Igualdad (no-discriminación): está presente en las legislaciones esenciales de derechos humanos de las Naciones Unidas; sin embargo, hay diferencias entre países de ley común y de ley civil. En el caso de ley común, las leyes se desarrollan por caso y contexto específicos. En ley civil se emplean leyes que cubren una larga lista de casos, basadas en principios comunes.

Igualdad formal (igualdad de trato): se basa en el principio de que elementos similares deberían ser tratados de manera similar; es decir, los individuos deberían tratarse en función de sus méritos propios y no de características irrelevantes (Barnard & Hepple, 2000).

Igualdad sustantiva (igualdad de resultados): los individuos son tratados diferente en función de sus diferencias, particularmente las desventajas asociadas con grupos protegidos, para lograr resultados justos en procesos de decisión. Las acciones afirmativas, y la reducción de la discriminación indirecta, se basan en un principio distributivo de justicia basado en la igualdad sustantiva (Barnard & Hepple, 2000).

Cuantificar el racismo es una tarea complicada. Un principio legal general es considerar la **infra-representación grupal** en la obtención de beneficios, o bien, la **sobre-representación grupal** en la negación de beneficios, como una medida cuantitativa de discriminación indirecta contra dicho grupo (el grupo protegido). También se concuerda en que las conclusiones estadísticas son una evidencia de discriminación a primera instancia, que puede refutarse con argumentos (las justificaciones legítimas ya mencionadas).

4.1 HERRAMIENTAS ESTADÍSTICAS

Como se mencionó en la sección anterior, ahora se plantearán conceptos generales relacionados con las herramientas estadísticas en el contexto del análisis de la discriminación. La referencia sigue estando dentro del mismo escrito, pero la sección es diferente. Como en la sección pasada, debe asumirse que la referencia para los conceptos presentados a continuación es una misma, hasta que comience la siguiente sección, correspondiente al contexto de la minería de datos. Dicha referencia es (Romei & Ruggieri, 2014, pp. 6–9). Nuevamente se incorporarán referencias adicionales cuando sea necesario.

Análisis estadístico de datos

La fuentes de información se clasifican de acuerdo al grado de control que tiene el analista. En las **experimentales** hay control sobre todas las variables independientes, en las **cuasi-experimentales** hay control sobre algunas de ellas, y en las **observacionales** no hay ningún control sobre estas variables.

Algunos elementos cruciales (especialmente porque las fuentes **observacionales** son más comunes) son: la solidez de la recolección de datos, la suficiencia de las variables relevantes, la mitigación de variables de confusión[11] y la inexistencia de explicaciones alternativas plausibles.

La estadística frecuentista se suele tomar como evidencia a primera instancia en casos legales. Una herramienta común para el análisis son las tablas de contingencia, que son tablas de dimensiones *k* x 2, para *k* grupos (pueden ser razas, edades, etc.), y denotan la frecuencia de ambos resultados posibles en cada grupo. Es decir, los resultados entre grupos se miden con una proporción de la gente en cada grupo con cada resultado. Estas proporciones son la representación estadística de la infra-representación grupal. El siguiente es un ejemplo para una población *n* compuesta por los grupos 1 y 2, y denota algunas cantidades de interés. Este ejemplo fue tomado de (Romei & Ruggieri, 2014, Fig. 1).

	beneficio		
grupo	**negado**	**concedido**	**total**
protegido	a	b	n_1
no protegido	c	d	n_2
total	m_1	m_2	n

Tabla 4.1. Tabla de contingencia de los grupos 1 y 2 en la población *n* (Romei & Ruggieri, 2014, Fig. 1).

$$p_1 = a/n_1 \qquad p_2 = c/n_2 \qquad p = m_1/n$$

$$RD = p_1 - p_2 \qquad RR = p_1/p_2 \qquad ED = p_1 - p \qquad ER = p_1 / p$$

$$RC = \frac{1-p_1}{1-p_2} \qquad OR = \frac{RR}{RC} = \frac{a/b}{c/d} = \frac{\frac{1-p_1}{1-p_2}}{\frac{p_1}{p_2}} = \frac{\frac{1-p_1}{p_1}}{\frac{1-p_2}{p_2}} \qquad EC = \frac{1-p_1}{1-p}$$

11 Una variable de confusión es aquella que al ignorarla o no controlarla, puede llevar a conclusiones erróneas en un estudio. Se trata de variables que son la verdadera razón detrás de resultados que el estudio asoció a otras causas (Skelly et al., 2012). Por ejemplo, un estudio tratando de encontrar la relación entre las ventas de diferentes modelos de tazas de té y sus tamaños, podría concluir que los modelos medianos tienen más ventas, y después descubrir que la verdadera causa fue que las tazas medianas tenían los diseños más populares. En ese caso, los diseños de las tazas serían variables de confusión.

Estas cantidades son medidas de la discriminación entre los grupos. Se pueden agrupar y definir de la siguiente manera:

Cantidades que comparan entre grupos:
RD: diferencia de riesgo o reducción absoluta de riesgo,
RR: radio de riesgo o riesgo relativo,
RC: posibilidad relativa o tasa de selección,
OR: radio de posibilidades.

Cantidades que comparan al grupo protegido con la población total:
ED: diferencia extendida,
ER: radio extendido o alzamiento extendido,
EC: posibilidad extendida.

Los autores extienden la discusión para hablar de las distribuciones que podrían tomar estas medidas, y de varios métodos involucrados con la aproximación de las mismas. Esto se desvía del enfoque de este trabajo por varias razones: se centra en la interpretación frecuentista de la probabilidad, las distribuciones para la probabilidad de no recibir el beneficio son distribuciones binomiales, lo cual es coherente con la perspectiva frecuentista, pero el planteamiento no es compatible con la perspectiva de un clasificador y su tarea de regresión. Estos detalles no serán relevantes para el presente estudio, ni siquiera para las descripciones de las implementaciones frecuentistas que se estudien, pero era necesario justificar su omisión. Estos detalles pueden consultarse en (Romei & Ruggieri, 2014, p. 7). Para concluir con este ejemplo, cabe mencionar que para evaluar si las diferencias entre grupos se deben a una tercera variable de control, debe incrementarse la dimensionalidad de la tabla acorde a los valores permitidos de dicha variable[12].

Se dice que el análisis estadístico mide la discriminación como un residuo: es la diferencia restante entre grupos, tras tomar en cuenta las demás influencias en el resultado, que dependen también de la pertenencia a los grupos (Quillian, 2008). Es difícil saber si ya se rindieron cuentas por todos los factores necesarios, por ello es más adecuado decir que este es un método para evaluar la proporción de una brecha grupal (racial, de género, etc.) justificable mediante los factores medidos, no para medir la discriminación tal cual. La sensibilidad estadística al sesgo causado por variables o factores omitidos también se ha estudiado[13].

La discriminación en un contexto de regresión es relevante para el presente trabajo. En la regresión lineal, el objetivo es explicar un resultado mediante la combinación lineal de variables independientes, gobernada por coeficientes para cada variable independiente, es decir:

$$Y_i = \beta_0 + \beta_1 X_{i,1} + ... + \beta_k X_{i,k} + \varepsilon_i$$

Donde el modelo (valores de los coeficientes $\boldsymbol{\beta}$) se aprende de los datos existentes[14].

12 Por ejemplo, si se estudian diferencias entre miembros de dos nacionalidades diferentes, una tercera variable podría ser la edad, para confirmar que las diferencias observadas no se deban a ésta. Ello implicará incrementar las dimensiones de la tabla para observar las proporciones no sólo para nacionalidades, sino también para grupos de edad (esto es el concepto de las variables de confusión descrito antes, en el contexto de una tabla de contingencia).

13 Nuevamente, no se aborda con detalle porque su enfoque se desvía hacia un área de la estadística frecuentista que no se discute en este trabajo, pues no trata directamente con clasificadores de aprendizaje computacional; sin embargo, se puede mencionar que algunos de estos métodos son: la prueba de chi cuadrada de Pearson, para tablas de contingencia; la prueba de rangos con signo de Wilcoxon, para el estudio de la diferencia de trato en casos con resultados ordinales; y para resultados continuos, la significación de la diferencia promedio de resultados se puede cuantificar con la prueba t de Student (Romei & Ruggieri, 2014, p. 8).

14 Por su parte la variable ε modela los residuales inherentes del modelo (la diferencia entre el resultado real y el que predice el modelo). Y el subíndice *i* indica que hay un conjunto de casos donde *i* = 1, ..., *n* (Romei & Ruggieri, 2014, p. 8).

Ahora bien, en una regresión lineal, la discriminación en la decisión puede denotarse como el peso dado a datos protegidos *D*, que a fin de cuentas son otras variables independientes para el modelo. Este peso es el coeficiente α en la regresión lineal, que multiplica a los datos protegidos *D*:

$$Y_i = \alpha D_i + \beta_0 + \beta_1 X_{i,1} + ... + \beta_k X_{i,k} + \varepsilon_i$$

En el caso de resultados binarios, el modelo adecuado es la regresión logística, que es la base de este trabajo[15]. α pasa a ser un coeficiente de variación del logaritmo de las posibilidades (*odds*) de Y_i, debida a la presencia de características discriminatorias, cuando las demás características o predictores se mantienen constantes.

Finalmente, en el caso del razonamiento causal, el efecto causal directo de características discriminatorias es la sensibilidad del resultado a variaciones en estas características, cuando las demás son constantes. La pregunta a contestar es: ¿el resultado del individuo hubiera sido el mismo manteniendo todas sus características invariantes, a excepción de las que son tanto irrelevantes como discriminatorias (raza, género, religión, etc.)? (FindLaw's United States Seventh Circuit Case and Opinions., s/f Carson versus Bethlehem Steel Corp.).

Análisis de datos con bases legales

Los objetivos científicos y los legales no son los mismos: existe un conflicto entre el rigor científico y la naturaleza del sistema legal, sin embargo, la evidencia estadística en casos de discriminación es persuasiva. En general, la estadística es una parte rutinaria y relevante en los procesos legales[16] (Gastwirth, 1992). Desde el punto de vista legal de la discriminación, el problema principal consiste en determinar la **población relevante** (o grupo de comparadores), **la medida de discriminación** que formalice la infra-representación[17] grupal, y el **umbral** que determine la evidencia de discriminación a primera instancia.

La población relevante es crucial en casos de discriminación indirecta porque influye la medida de la infra-representación. Para defenderse de alegaciones de discriminación con bases estadísticas es común apelar a la paradoja de Simpson[18], fundamentada precisamente en la elección de la población relevante.

En el ejemplo presentado con la tabla de contingencia, se ilustraron varias medidas de discriminación. La medida empleada varía entre autoridades legales: en Reino Unido se usa la diferencia de riesgo (RD), en la Unión Europea se trata el radio de riesgo (RR), y en los Estados Unidos se enfocan en la tasa de selección (RC). En investigaciones legales se usa ampliamente el radio de posibilidades (OR). Estas medidas son sólo aproximaciones de la realidad, y como tal, la selección de una medida, y los métodos para su evaluación, pueden ser parte del debate en un juicio.

15 Y como tal, los detalles de su uso en el análisis de la discriminación, y del modelo mismo, serán tratados con profundidad en capítulos posteriores.

16 Gastwirth (1992) presenta un análisis interesante del tema, incluyendo el conflicto mencionado y el rol de la estadística en los procesos legales. Propone que el objetivo científico, el entendimiento de los mecanismos que generan los datos, tiene el fin de predecir resultados nuevos, mientras que en la ley el objetivo es resolver disputas particulares. Esto implica que los estándares generales en el contexto científico no aplican en un contexto legal, donde hay estándares particulares para distintos casos. Es una discusión interesante, pero queda fuera del enfoque de este trabajo.

17 La infra-representación puede tener un nombre engañoso. No se refiere a que en la muestra considerada haya menos elementos en total del grupo protegido, sino a que la proporción que recibe el resultado beneficioso para dicho grupo, es menor a la proporción que lo recibe en el otro grupo.

18 Esta paradoja consiste en la inversión de la asociación o tendencia entre dos variables. La relación entre ambas puede parecer consistente a través de diversas poblaciones, pero al considerar el agregado de las mismas en una población total, la relación se invierte (lo mismo puede pasar para una relación que parece consistente en una población total, y cambia al considerarla como un grupo de poblaciones individuales). Más detalles e ilustraciones en (Malinas & Bigelow, 2016).

En cuanto a umbrales de infra-representación, en la Corte de Justicia Europea se ha enfatizado que para considerarse como evidencia estadística, la cifra debe ser "sustancial", pero no hay un umbral estricto en la Unión Europea. En los Estados Unidos hay dos umbrales formales: la regla de los cuatro quintos[19] y la regla de Castaneda (o Castañeda)[20].

4.2 ANÁLISIS DISCRIMINATORIO DENTRO DEL CONTEXTO DE MINERÍA DE DATOS Y DESCUBRIMIENTO DE CONOCIMIENTO[21]

Para la última parte de este tema, se describe el análisis de la discriminación en un contexto de minería de datos, aplicable y relevante para el aprendizaje computacional en general. La referencia central seguirá siendo la misma, pero en una nueva sección: (Romei & Ruggieri, 2014, pp. 38–40). Las referencias adicionales se presentan en donde se requiera. En este contexto, se trata tanto el **descubrimiento** como la **prevención** de la discriminación.

Descubrimiento:

Consiste en hallar situaciones discriminatorias en archivos de decisiones tomadas. No es una prueba de hipótesis estadística, sino un esfuerzo por descubrir contextos de posible discriminación. Para este objetivo, se ha propuesto la extracción de reglas **potencialmente discriminatorias (PD)** de los conjuntos de datos (*datasets*) (Ruggieri et al., 2010), (Pedreschi et al., 2008). Estas reglas son de la forma **A, B → C**, donde **A** denota un grupo protegido, **B** algún contexto donde se investiga la posible discriminación, y **C** es el resultado[22]. (por ejemplo, **A** es alguna raza, **B** solicitar un puesto administrativo, y **C** recibir empleo). La posible sobre-representación para un resultado negativo **C** y un contexto **B** se cuantifica con el radio extendido (ER)[23]. Acorde a este valor se jerarquizan las reglas PD.

Hay dos problemas notables con este método: primero, las métricas de discriminación mezclan decisiones correspondientes a individuos que pueden ser muy diferentes en factores cruciales, resultando en un exceso de reglas a examinar para descartar la posibilidad de una discriminación justificable. Y en segundo lugar, el resultado será un conjunto de reglas de clasificación que proveen nichos locales y traslapados de posible discriminación: el método carece de una descripción global que muestre quiénes sufren la discriminación (y quiénes no) individualmente.

Otra propuesta interesante es la **prueba de situaciones** (Luong et al., 2011). Consiste en ubicar a cada miembro del grupo protegido que haya recibido un resultado negativo con base en sus características (relevantes para el resultado), y después, con base en una función de distancia para cuantificar la similitud entre individuos, hallar sujetos prueba cercanos o similares provenientes de

19 RC < 0.8 significa discriminación bajo este umbral (*Code of Federal Regulations*, s/f). Más detalles y discusión con respecto a su uso y definición pueden hallarse en (Bobko & Roth, 2004).

20 El número de elementos seleccionados de un grupo protegido no debe estar más de tres desviaciones estándar por debajo de la media (que es el número de elementos esperado en una selección aleatoria) (Sugrue & Fairley, s/f). Esta norma se derivó de un caso jurídico real, en particular, del proceso de selección de jurado (*Castaneda v. Partida, 1977*, s/f).

21 Es bueno tener presente la distinción entre minería de datos, descubrimiento de conocimiento, y aprendizaje computacional: el descubrimiento de conocimiento se puede interpretar como uno de los dos tipos principales de aprendizaje computacional: el aprendizaje no supervisado (Murphy, 2012, p. 2). Y la minería de datos es un paso dentro del descubrimiento de conocimiento (*KDD Process/Overview*, s/f).

22 Dicho resultado puede ser positivo o negativo, pero no es una variable. Es decir, **C** podría representar recibir un empleo o ser rechazado para el mismo, pero se tiene que adherir a alguno de los dos casos, no ser una variable que puede dar cualquiera de los dos resultados. Si **C** es positivo, se estudia la infra-representación del grupo protegido, si es negativo, se estudia la sobre-representación (Romei & Ruggieri, 2014, p. 38).

23 Podrían usarse otras medidas, lo cual fue estudiado en (Pedreschi et al., 2012), donde se concluye que la medida seleccionada afecta críticamente la jerarquía obtenida de las reglas PD, y en particular, las mayores diferencias en la jerarquía se observan al pasar de usar el riesgo relativo (RR), a usar la tasa de selección (RC) como medida de referencia.

otros grupos, para así determinar el grado de discriminación mediante las métricas usuales (RR, RC, ER, etc.) comparando sólo con los individuos similares, no con cualquiera del grupo ajeno. Esto da como resultado una descripción global de quiénes han sufrido discriminación, que se puede tener en forma de una tarea estándar de clasificación.

No siempre habrá atributos en los datos que muestren la pertenencia a grupos protegidos. Una propuesta es inferir las reglas para lidiar con el descubrimiento indirecto[24] de la discriminación (Ruggieri et al., 2010). Por ejemplo, para la regla **A, B → C**, si no se dispone de **A**, de todos modos se puede deducir con reglas de clasificación sin elementos protegidos, en la forma **D, B → C**, donde **D** es una característica cuya correlación con **A** es fuerte. La información referente a este tipo de correlaciones denota conocimiento previo, y es en sí una regla asociativa.

Prevención:

Esta tarea consiste en extraer modelos (clasificadores) que hallen un balance entre precisión y no-discriminación. Estos modelos, basándose en la información histórica, podrían "heredar" prejuicios y discriminación estadística.

Un enfoque ingenuo es tratar de erradicar los atributos protegidos y entrenar de esta forma a los clasificadores: el modelo aún podrá aprender decisiones discriminatorias mediante otros atributos fuertemente correlacionados a los protegidos, estos atributos alternativos funcionan como "proxies"[25] para el modelo. Una de las primeras propuestas trató con modelos de regresión logística en el área de puntajes de crédito[26], y consiste en modificar los pesos de los coeficientes de regresión obtenidos (Fortowsky & LaCour-Little, 2001).

En la minería de datos, hay cuatro estrategias para mitigar la discriminación. Éstas no se excluyen mutuamente, y pueden considerarse en el contexto específico del aprendizaje computacional:

Distorsión controlada de los datos de entrenamiento: su enfoque está en el preprocesado de datos, y básicamente consiste en modificar o perturbar los datos para eliminar la discriminación presente, para así poder usarlos en la construcción del clasificador sin que éste herede los prejuicios de los datos (Kamiran & Calders, 2012). La gran ventaja de esta estrategia es que los métodos desarrollados son independientes del modelo y algoritmo de aprendizaje empleados, pues tratan únicamente con los datos.

Modificar el algoritmo de aprendizaje: esta estrategia se enfoca en el proceso interno del clasificador (*in-processing*), y se desarrolla de manera integrada con los criterios de anti-discriminación. El objetivo es modificar el algoritmo del clasificador para que la salida tome en cuenta los criterios anti-discriminatorios, esto involucra que el clasificador no tenga el objetivo único de realizar predicciones precisas con base en los datos observados, sino que considere también la discriminación que podría darse en estas predicciones. En este caso los métodos son dependientes del modelo, ya que distintos clasificadores tienen distintos parámetros y estructuras.

Post-procesado del clasificador: en esta estrategia el clasificador nuevamente es modificado, pero no desde el algoritmo de aprendizaje, sino en los parámetros del

24 Nótese que el descubrimiento es el que es indirecto, la discriminación descubierta puede ser directa o indirecta.

25 Este término hace alusión a que desde el punto de vista del modelo, la parte de las decisiones que se hubiera basado injustamente en atributos protegidos, se basará en estas otras variables, llevando al mismo resultado aunque se hayan ignorado las características discriminatorias.

26 Puntajes calculados y usados por prestadores para determinar si un candidato es aceptable para recibir el préstamo (*Puntajes de crédito*, 2013).

clasificador obtenido. Los métodos de esta estrategia también son dependientes del tipo de modelo usado.

Corrección de proporcionalidad: los métodos para esta última estrategia no hacen ninguna modificación al clasificador ni a los datos, sino que corrigen las predicciones obtenidas del mismo, buscando mantener la proporcionalidad de decisiones entre los distintos grupos. Una propuesta es corregir predicciones cerca de la frontera de decisión de clasificadores probabilísticos, pues son decisiones propensas a verse afectadas por la discriminación estadística (Kamiran & Calders, 2012).

Los conceptos de este capítulo cubren bastante más que los fundamentos teóricos necesarios para el enfoque, las implementaciones y los resultados presentados en este escrito. Sin embargo, es importante desarrollar una revisión general del tema para ilustrar con claridad el alcance, las limitaciones y el contexto de este trabajo.

Para cerrar este capítulo, parece interesante y adecuado citar la conclusión de la referencia central del mismo. Los autores afirman que el interés en el tema ha sido estimulado recientemente por la llegada de la "sociedad del conocimiento" (*knowledge society*): una sociedad donde las decisiones se pueden tomar con base en modelos, patrones, perfiles y reglas del comportamiento humano, obtenidas de las trazas digitales generadas como un efecto secundario de nuestra vida y su implícita interacción en línea (Romei & Ruggieri, 2014, p. 40).

REFERENCIAS

Barnard, C., & Hepple, B. (2000). Substantive Equality. *The Cambridge Law Journal*, *59*(3), 562–585. JSTOR.

Bobko, P., & Roth, P. (2004). THE FOUR-FIFTHS RULE FOR ASSESSING ADVERSE IMPACT: AN ARITHMETIC, INTUITIVE, AND LOGICAL ANALYSIS OF THE RULE AND IMPLICATIONS FOR FUTURE RESEARCH AND PRACTICE. *Research in Personnel and Human Resources Management*, *23*, 177–198. https://doi.org/10.1016/S0742-7301(04)23004-3

Castaneda v. Partida, 1977. (s/f). Recuperado el 9 de octubre de 2019, de http://foofus.net/goons/foofus/lawSchool/constitutionalLawII/Castaneda_v_Partida.html

Code of Federal Regulations. (s/f). Recuperado el 21 de mayo de 2020, de https://www.govinfo.gov/content/pkg/CFR-2011-title29-vol4/xml/CFR-2011-title29-vol4-part1607.xml

Ellis, E. (2005). *EU Anti-discrimination Law*. Oxford University Press. https://books.google.se/books?id=1FJ2QgAACAAJ

FindLaw's United States Seventh Circuit case and opinions. (s/f). Findlaw. Recuperado el 21 de mayo de 2020, de https://caselaw.findlaw.com/us-7th-circuit/1304532.html

Fortowsky, E., & LaCour-Little, M. (2001). *Credit Scoring and Disparate Impact*. 1–23.

Gastwirth, J. (1992). Statistical Reasoning in the Legal Setting. *American Statistician - AMER STATIST*, *46*, 55–69. https://doi.org/10.1080/00031305.1992.10475851

Harford, T. (2008). *The Logic of Life: The Rational Economics of an Irrational World*. Random House Publishing Group. https://books.google.se/books?id=ugWnyvANaEsC

Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, *33*(1), 1–33. https://doi.org/10.1007/s10115-011-0463-8

Kanter, R. M. (1977). Some Effects of Proportions on Group Life: Skewed Sex Ratios and Responses to Token Women. *American Journal of Sociology*, *82*(5), 965–990.

KDD Process/Overview. (s/f). Recuperado el 28 de mayo de 2020, de http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1_kdd.html

Luong, B. T., Ruggieri, S., & Turini, F. (2011). K-NN as an implementation of situation testing for discrimination discovery and prevention. *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '11*, 502. https://doi.org/10.1145/2020408.2020488

Malinas, G., & Bigelow, J. (2016). Simpson's Paradox. En E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2016). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2016/entries/paradox-simpson/

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT press.

Newman, D. (2014). Sociology: Exploring the architecture of everyday life / David M. Newman. *SERBIULA (sistema Librum 2.0)*.

Pedreschi, D., Ruggieri, S., & Turini, F. (2012). A Study of Top-K Measures for Discrimination Discovery. *Proceedings of the ACM Symposium on Applied Computing*. https://doi.org/10.1145/2245276.2245303

Pedreschi, D., Ruggieri, S., & Turini, F. (2008). Discrimination-aware data mining. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 560–568. https://doi.org/10.1145/1401890.1401959

Puntajes de crédito. (2013, septiembre 1). Información para consumidores. https://www.consumidor.ftc.gov/articulos/s0152-puntajes-de-credito

Quillian, L. (2008). New Approaches to Understanding Racial Prejudice and Discrimination. *Annual Review of Sociology*, *32*. https://doi.org/10.1146/annurev.soc.32.061604.123132

Romei, A., & Ruggieri, S. (2014). A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review*, *29*(5), 582–638.

Ruggieri, S., Pedreschi, D., & Turini, F. (2010). Data Mining for Discrimination Discovery. *TKDD*, *4*. https://doi.org/10.1145/1754428.1754432

Skelly, A. C., Dettori, J. R., & Brodt, E. D. (2012). Assessing bias: The importance of considering confounding. *Evidence-Based Spine-Care Journal*, *3*(1), 9–12. PubMed. https://doi.org/10.1055/s-0031-1298595

Sowell, T. (2004). *Affirmative Action Around the World*. Yale University Press; JSTOR. www.jstor.org/stable/j.ctt1npfgb

Sugrue, T. J., & Fairley, W. B. (s/f). *A Case of Unexamined Assumptions: The Use and Misuse of the Statistical Analysis of Castaneda/Hazelwood in Discrimination Litigation*. *24*, 38.

The United Nations. (1966). International Convention on the Elimination of All Forms of Racial Discrimination. *Treaty Series*, *660*, 195.

Tobler, C. (2008). *Limits and potential of the concept of indirect discrimination*.

5. CLASIFICADORES

En este capítulo se describen los detalles de los modelos usados en este trabajo. Las implementaciones usarán modelos de regresión logística, pero en este capítulo también se define la tarea de clasificación y se revisan otros dos clasificadores importantes, para ilustrar el concepto general, y para hacer una aclaración crucial en cuanto al término de "clasificador bayesiano".

Para comenzar, se describirá la tarea de clasificación en el contexto general del aprendizaje computacional. De acuerdo a Murphy (2012, pp. 2, 8-9) se puede hablar de dos tipos principales de aprendizaje: **supervisado** y **no supervisado**[27]; en la práctica, el aprendizaje supervisado es el más común de los dos, y el desarrollo del presente escrito entra en esta categoría. A su vez, las dos tareas esenciales de este tipo de aprendizaje son **clasificación** y **regresión**.

Este capítulo y el trabajo mismo se enfocan en la clasificación. El objetivo es aprender una relación o mapeo entre entrada (usualmente varias entradas o una entrada multidimensional) y salida, donde la salida es una respuesta discreta que representa dos o más clases. Es decir, a partir de los valores de ciertos atributos de entrada para algún elemento, se le asigna una clase[28]. Los clasificadores pueden tener múltiples salidas[29], pero lo más común es que tengan una sola salida, que representará una de varias clases.

El problema se formaliza con la suposición de que la salida y es función de las entradas $\mathbf{x}$, es decir: $y = f(\mathbf{x})$. Entonces el problema es aproximar la función desconocida f. Esto se hace con un conjunto de entrenamiento (*training set*) ya clasificado o etiquetado, es decir, un conjunto de datos ordenados en registros, donde cada registro tiene una entrada $\mathbf{x}$ con su correspondiente salida y[30]. Aprendiendo de estos datos, surge entonces la función aproximada: $\hat{y} = \hat{f}(\boldsymbol{x})$. Y con dicha función aproximada, se pueden hacer predicciones para nuevas entradas $\mathbf{x}^*$ cuyas salidas y^* se desconozcan. El objetivo es lograr la **generalización**, que consiste en obtener una función aproximada que estime efectivamente las salidas para nuevos datos de entrada (la precisión dentro del *training set* es irrelevante desde este punto de vista, pues para esos datos ya se conoce la salida).

Por su parte, la tarea de regresión atiende el mismo problema, pero la respuesta o salida y no representa clases discretas, sino que es de carácter continuo[31]. A continuación se describirán en términos generales, varios tipos de clasificadores, donde destaca la regresión logística, que será el método utilizado en el desarrollo de este trabajo.

5.1 CLASIFICADORES BAYESIANOS[32]

La **regla de clasificación bayesiana** consiste en asignar una entrada o patrón desconocido $\mathbf{x}$ a alguna clase y_i, para la cual la probabilidad a posteriori es la máxima (con relación a las otras clases). Es decir, considerando M clases, un patrón $\mathbf{x}$ se asigna a la clase y_i cuando:

$$y_i = \arg\max_{y_j} \mathrm{P}(y_j|\mathbf{x}) \; ; \quad j = 1, 2, \ldots, \mathrm{M}.$$

Lo cual significa que la clase asignada y_i es la que maximiza la probabilidad a posteriori $\mathrm{P}(y_j|\mathbf{x})$ de entre todas las clases y_j. Antes de recibir cualquier observación, la incertidumbre y conocimiento

27 Hay un tercer tipo, el **aprendizaje reforzado** (*reinforcement learning*) (Murphy, 2012, p. 2).

28 Un ejemplo sería clasificar a una persona como "alta" o "baja", a partir de su atributo "estatura".

29 Las personas podrían clasificarse como "altas" y "bajas", pero también como "viejos" y "jóvenes", teniendo entonces una salida total compuesta de dos salidas binarias. Los componentes de la salida total no requieren ser forzosamente binarios.

30 Ya que cada registro tiene un vector de entradas $\mathbf{x}$ con una salida y, es común hacer referencia a la totalidad de entradas como una matriz $\mathbf{X}$, y a las salidas como un vector $\mathbf{y}$.

31 Por ejemplo, un problema de regresión sería predecir la temperatura de una habitación, en función de entradas como la hora, el clima, etc. En este trabajo el enfoque está en problemas de clasificación.

32 (Theodoridis, 2015, pp. 276-280).

con respecto al proceso se reflejan con las probabilidades a priori P(y_i) de las mismas clases. Es hasta después de recibir las observaciones que el conocimiento se "actualiza", y entonces es representado por la probabilidad a posteriori, de acuerdo al teorema de Bayes:

$$\mathrm{P}(y_j|\mathbf{x}) = \frac{p(\mathbf{x}|y_j)\mathrm{P}(y_j)}{p(\mathbf{x})} \; ; \quad j=1, 2, ..., \mathrm{M}.$$

Donde las probabilidades de la forma $p(\mathbf{x}|y_i)$ son las densidades (PDFs) condicionales en cada clase. La densidad en el denominador $p(\mathbf{x})$ es independiente de las clases, no afecta la tarea de maximizar la probabilidad a posteriori, y entonces la regla de clasificación se convierte en:

$$y_i = \arg\max_{y_i} p(\mathbf{x}|y_j)\mathrm{P}(y_j) \; ; \quad j=1,2,..,\mathrm{M}.$$

Un detalle de los clasificadores bayesianos es que reducen el **error de clasificación**; dicho error es la probabilidad de que a una entrada se le asigne una clase equivocada, dada la clase a la que pertenece. Al diseñar un clasificador, el objetivo es dividir el espacio de las entradas, definiendo clases que correspondan a regiones de dicho espacio. La probabilidad de que el clasificador asigne una clase a una entrada, cuando pertenece a otra, es el error de clasificación. Los clasificadores bayesianos minimizan esta probabilidad con respecto a todas las regiones consideradas como clases.

El **riesgo promedio** es otro elemento relacionado con clasificadores bayesianos: el error de clasificación no siempre es la función que se requiere minimizar, pues hay varios contextos dentro de los cuales algunos errores son más costosos que otros, y el error de clasificación es simétrico para ambos tipos de errores, así que no tomaría este hecho en cuenta. El riesgo promedio es una función de costo basada en el error de clasificación, que asigna pesos relativos a cada tipo de error para darles prioridades diferentes.[33]

La **opción de rechazo** es otro elemento relacionado con aplicaciones donde las decisiones son delicadas, y se refiere a establecer cierto umbral para clasificar elementos. Es decir, si hay una o más clases que tengan una probabilidad lo suficientemente cercana a la clase de máxima probabilidad, se puede no clasificar la entrada en cuestión ya que hay varias clases que serían razonables para ésta. La especificación matemática para la opción de rechazo es arbitraria, y depende de la aplicación.

Clasificadores bayesianos y metodología bayesiana

Una distinción crucial para este trabajo es entre estos "clasificadores bayesianos" y la metodología bayesiana. Los "clasificadores bayesianos" de esta sección emplean el teorema de Bayes para estimar los parámetros necesarios y realizan inferencias sin ningún tipo de incertidumbre sobre el modelo; la única incertidumbre en ellos es la que se da en la determinación de las probabilidades empleadas para clasificar, pero el proceso no tiene una incertidumbre sobre sus parámetros en el sentido al que se refiere una metodología bayesiana: una vez que se determina la opción de mayor probabilidad, se selecciona sin más, o en su defecto se elige la opción de rechazo.[34] A pesar de esto,

33 Asignar más peso a algunos errores sobre otros no es exclusivo de estos clasificadores. En general la expresión matemática de los clasificadores que determina el costo o peso de los errores, puede ser alterada para darle más peso a tipos de errores específicos. Un ejemplo de esto se encuentra en (*Choosing Logisitic Regression's Cutoff Value for Unbalanced Dataset*, s. f.).

34 Un ejemplo numérico puede ayudar a ilustrar este concepto: si el clasificador asigna una probabilidad de 0.8 a una clase y 0.2 a otra, seleccionará la primera (por ahora se ignora la opción de rechazo, que de todos modos sería inusual para un caso tan contrastante), sin ninguna información acerca de cuán seguro está de esos valores de 0.8 y 0.2 que fueron calculados con el teorema de Bayes.

son un ejemplo útil que ilustra la aplicación del teorema de Bayes (*MLPR w3a - Machine Learning and Pattern Recognition*, s. f.).

En la metodología bayesiana se pueden usar otros clasificadores además de los "clasificadores bayesianos" descritos en esta sección (incluso es algo común), y la característica fundamental que los identifica como métodos bayesianos es la incertidumbre sobre los parámetros y su tratamiento mediante distribuciones de probabilidad en vez de valores puntuales, independientemente de que los mismos sean estimados con el teorema de Bayes como en los "clasificadores bayesianos", o por otros métodos, como se da en la regresión logística (*MLPR w3a - Machine Learning and Pattern Recognition*, s. f.).

5.2 KNN[35]

Los clasificadores *k-nearest neighbor*s (k vecinos más próximos o KNN por las siglas en inglés) son modelos no paramétricos, sencillos en concepto, con un solo parámetro, *K*[36], que es un número natural. El funcionamiento es simple: para cada nueva entrada, se observan los *K* vecinos más cercanos y se le asigna la clase a la cual pertenezca la mayoría de ellos. Inicialmente, este cálculo se hace con base en los puntos de entrenamiento que ya se tienen disponibles, pero conforme vienen nuevos datos, éstos también deben tomarse en cuenta para las próximas decisiones ya que se convierten en nuevos "vecinos". Dadas M clases, *K* no debería ser múltiplo de M, pues esto podría resultar en empates.

KNN es mucho más sencillo que los clasificadores bayesianos, y tiende a comportarse como éstos cuando el número de datos o registros N y el parámetro *K* tienden al infinito, y si además la relación *K*/N tiende a cero. También cabe mencionar que el error de clasificación usando KNN nunca excederá al doble del error de clasificación bayesiano cuando *K* es 1, y conforme aumente *K* ese límite superior se reduce.

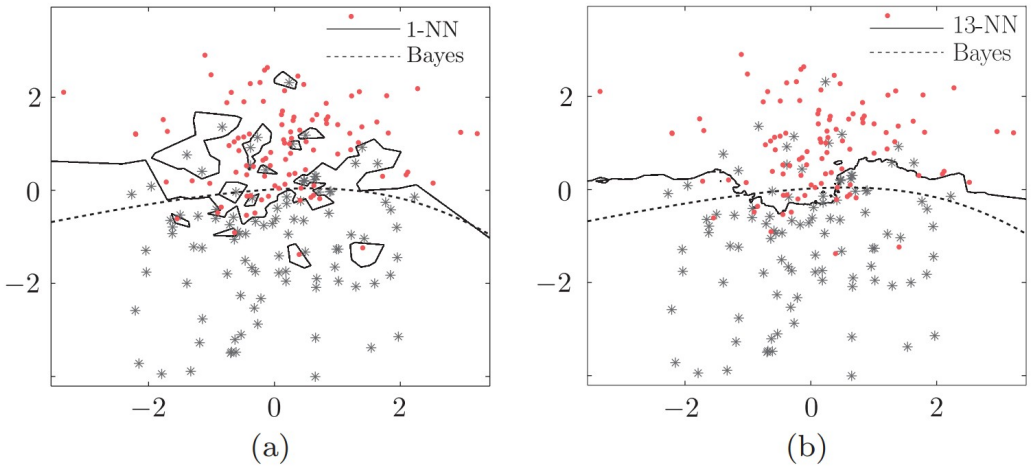


Figura 5.1. Tareas de clasificación comparando un clasificador KNN con un clasificador bayesiano, para el caso con un parámetro unitario *K* (a), y para el caso con *K* = 13 (b). Con el aumento de *K*, la frontera es más suave y se aproxima a la del clasificador bayesiano (Theodoridis, 2015, Fig. 7.8).

35 (Theodoridis, 2015, pp. 288-290).

36 Se dice que es no paramétrico porque no asume nada acerca de la distribución de los datos a clasificar, no porque no tenga ningún tipo de parámetro. De hecho, en los modelos no paramétricos el número de parámetros crece conforme lo hace la cantidad de datos (Murphy, 2012, p. 16).

5.3 REGRESIÓN LOGÍSTICA[37]

En términos probabilísticos, los clasificadores mencionados hasta ahora generan modelos de la distribución conjunta $p(y, \mathbf{x})$, y tras hacerla condicional en **x**, se deriva la distribución a posteriori para hacer inferencias: $p(y|\mathbf{x})$. A este método se le llama **generativo**. La otra opción, el método **discriminativo**, consiste en modelar directamente $p(y|\mathbf{x})$. La regresión logística es un clasificador del segundo tipo (Theodoridis, 2015, pp. 63-64), (Murphy, 2012, pp. 267-271).

Este clasificador *binario* corresponde a la expresión:

$$p(y|\mathbf{x}, \mathbf{w}) = Ber(y|sigm(\mathbf{w}^T\mathbf{x}))$$

Esto quiere decir que la distribución de probabilidad a posteriori para la salida *y* se modela como una distribución de Bernoulli: el único parámetro de la distribución de Bernoulli es la probabilidad de éxito, normalmente denominada como "θ" (Murphy, 2012, pp. 34-35). En este caso, dicho parámetro es modelado mediante la función sigmoide de la combinación lineal de componentes de entrada, con sus respectivos pesos. El vector **w** representa los pesos correspondientes a cada dimensión de entrada, y el vector **x** es la entrada, por lo que el producto del vector transpuesto de los pesos por el vector de entrada, da la combinación lineal de cada componente de entrada. Por ejemplo, con entradas bidimensionales:

$$\mathbf{w} = [w_1, w_2]^T \quad \mathbf{x} = [x_1, x_2]^T \qquad \mathbf{w}^T\mathbf{x} = w_1x_1 + w_2x_2$$

Se omitió el término de sesgo o *bias* en la combinación resultante (w_0 o b), que es un término "constante" en el sentido de que no se multiplica por componentes del vector de entrada, o bien, es como si el componente de la primera dimensión de todo vector de entrada **x** fuese siempre la unidad. Este término también forma parte de la combinación lineal a la que se le aplica la función sigmoide.

La función sigmoide tiene la cualidad de limitar su rango entre los valores 0 y 1, y por ello se usa para modelar el parámetro de "probabilidad de éxito" de la distribución de Bernoulli. Esta función provoca que la relación entre entrada y salida de la regresión logística sea no lineal. La función sigmoide se define de la siguiente manera:

$$sigm(a) = \frac{1}{1+e^{-a}}$$

Es así que la regresión logística modela la "probabilidad de éxito"[38] del modelo de Bernoulli como la función sigmoide de la combinación lineal de las covariables o entradas **X** (especificada por los parámetros **w** del modelo, incluido el término de sesgo *b*). Dependiendo de cómo sea la entrada, se tendrán distintas probabilidades de éxito, definidas por $sigm(\mathbf{w}^T\mathbf{x_i} + b)$ para cada elemento *i*.

Vale la pena hablar un poco más de la frontera entre clases que define una función sigmoide, pues será central en el ejemplo que se desarrolla sobre datos sintéticos más adelante. A continuación se muestra una función sigmoide en una dimensión (que correspondería a tener una sola covariable de entrada, y por ende, un único parámetro a variar en la regresión logística[39]), donde la frontera es el punto x = 0, pues es donde hay una probabilidad equivalente de pertenecer a cualquier clase (el valor de la función sigmoide es 0.5 en este punto):

---

37 (Murphy, 2012, pp. 245-253), (*MLPR w3c - Machine Learning and Pattern Recognition*, s. f.).

38 En un contexto de regresión logística, el "éxito" y el "fracaso", que son los dos resultados posibles en una distribución de Bernoulli, son las dos clases consideradas en la tarea de clasificación. No necesariamente tienen connotaciones positiva y negativa.

39 Adicionalmente, en la figura no se considera el valor del único parámetro *w* que debería multiplicar a la entrada x, o bien, se considera unitario. Tampoco se considera el término de sesgo *b*, que se suma como una constante a la entrada.

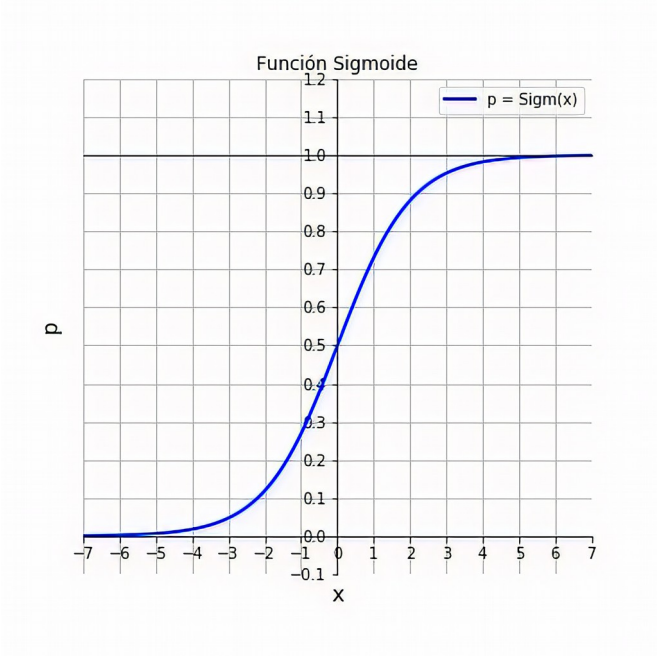


Figura 5.2. Función sigmoide o sigmoidea.

Conforme los puntos o entradas x se alejan de la frontera, las probabilidades de pertenecer a las clases tienden a 1 y 0 (dependiendo de qué lado de la frontera esté la entrada). Conforme aumentan las covariables, aumentan los parámetros y las dimensiones del espacio que la función sigmoide divide: con dos covariables (como en el ejemplo de datos sintéticos desarrollado más adelante) la frontera ya no es un punto, sino una línea; con tres, es un plano, y con más de tres son hiper-planos n-dimensionales. Al considerar la frontera de la regresión logística, se debe tener en cuenta que sólo representa la región en el espacio de las covariables donde se tiene la misma probabilidad de pertenecer a cualquiera de las dos clases, mas no representa la relación de las covariables con la salida, que es la función sigmoide representada en una dimensión superior a la de la frontera: en el ejemplo presentado aquí, el espacio de las covariables es la recta de los números reales, la frontera es el punto x = 0, y la relación entre las covariables y la salida es la función sigmoide, que se representa en el plano.

Hay una gran variedad de clasificadores, pero en este capítulo el objetivo es introducir la herramienta principal: la regresión logística, y también hablar de los "clasificadores bayesianos" para aclarar su distinción de la metodología bayesiana en general. También se discute el clasificador KNN, porque es intuitivo, fácil de explicar y un buen referente para mostrar el funcionamiento de los clasificadores generativos y la tarea de clasificación en general.

5.4 REGRESIÓN LOGÍSTICA BAYESIANA[40]

En esta sección se describirán las consideraciones básicas para un clasificador bayesiano, con el significado que tiene en el título de este trabajo. La idea es ilustrar cuáles son los cambios con respecto a un modelo no bayesiano de regresión logística, pero más que cambios, son consideraciones adicionales que deben tomarse en cuenta para el método bayesiano, pues el modelo sigue siendo de regresión logística, y como tal, conserva la estructura básica desarrollada en la sección anterior.

40 (Murphy, 2012, pp. 254-261).

Obtener la distribución a posteriori sobre los parámetros de una regresión logística no puede lograrse con exactitud, ya que no hay ninguna distribución a priori conjugada que pueda emplearse con este fin. Es entonces necesario obtener una aproximación. Hay varios métodos que tienen el propósito de aproximar la distribución a posteriori en un contexto general, no sólo aplicables para la regresión logística.

Aproximación de Laplace:

Consiste en aproximar la distribución a posteriori con una distribución normal. Es un método relativamente simple, y se basa en utilizar la curvatura de la distribución verdadera para hallar una aproximación normal adecuada. Se construye para tener la misma moda que la distribución verdadera. Al ser una distribución normal, será simétrica, lo cual difícilmente será el caso para la distribución verdadera, pero al menos puede proveer alguna información con respecto a la incertidumbre (Murphy, 2012, pp. 255-257), (MacKay, 2003, pp. 341-342).

Inferencia Variacional:

Este es el método que se emplea en los ejemplos de los siguientes capítulos. Es muy útil porque hay muchos fenómenos en los que una aproximación con la distribución normal es irrazonable, y la inferencia variacional es un método más general para aproximar distribuciones intratables analíticamente. En este método se selecciona una aproximación $q(\mathbf{w})$ para la distribución a posteriori verdadera $p(\mathbf{w} \mid \mathbf{D})$[41], y después se ajustan sus parámetros para que sea tan cercana a $p(\mathbf{w} \mid \mathbf{D})$ como sea posible. La diferencia entre las distribuciones se convierte entonces en la función a minimizar, y la inferencia se convierte en un problema de optimización, donde hay libertad para decidir el balance entre rapidez y precisión (como en cualquier otra optimización). Es un método mucho más flexible, pues para la distribución seleccionada $q(\mathbf{w})$ hay otras opciones diferentes a la distribución normal (Murphy, 2012, pp. 731-732).

La **divergencia de Kullback-Leibler**, **divergencia KL** o **entropía relativa**, es una función que determina la "distancia" entre dos distribuciones probabilísticas, y en este contexto es una función ideal para minimizar. Es importante notar que $\mathrm{KL}(p^* \parallel q)$ no es igual a $\mathrm{KL}(q \parallel p^*)$[42], así que es relevante decidir cuál planteamiento de esta divergencia se minimizará. Ya que se conoce $q(\mathbf{w})$ y no $p(\mathbf{w} \mid \mathbf{D})$, la elección siempre es $\mathrm{KL}(q \parallel p^*)$, ya que esta forma implica obtener valores esperados con respecto a la distribución de $q(\mathbf{w})$, lo cual no sería posible para la distribución desconocida e intratable $p(\mathbf{w} \mid \mathbf{D})$ (Murphy, 2012, pp. 57-58), (Murphy, 2012, pp. 733-734).

Aún con los ajustes realizados, es necesario emplear una distribución no normalizada $p_1(\mathbf{w})$, ya que $p(\mathbf{w} \mid \mathbf{D})$, la distribución a posteriori, es por definición una distribución condicional dados los datos observados (por eso es "a posteriori"), y por lo general la constante de normalización no puede obtenerse analíticamente (que de hecho es la distribución de los datos $p(\mathbf{D})$, la cual no es necesaria para realizar predicciones de lo que suceda con nuevos datos). Reemplazar $p(\mathbf{w} \mid \mathbf{D})$ por $p_1(\mathbf{w})$ lleva a una nueva función objetivo para minimizar.

Al ser $p(\mathbf{D})$ la constante de normalización, la distribución no normalizada $p_1(\mathbf{w})$ no es mas que el producto de dicha constante por la distribución normalizada $p(\mathbf{w} \mid \mathbf{D})$, y por la regla del producto de la probabilidad, se tiene entonces: $p_1(\mathbf{w}) = p(\mathbf{w} \mid \mathbf{D})p(\mathbf{D}) = p(\mathbf{w}, \mathbf{D}) = p(\mathbf{D} \mid \mathbf{w})p(\mathbf{w})$. En la última forma de la expresión se puede ver entonces que $p_1(\mathbf{w})$ no es más que el producto de $p(\mathbf{D} \mid \mathbf{w})$, que es la **verosimilitud** de los parámetros, y $p(\mathbf{w})$, que es la distribución a priori sobre los

41 Nótese que aquí los datos se expresan como **D**, y no como entradas **X** y salidas **y**. Los parámetros pueden encontrarse como **x** en la literatura, pero para evitar confusiones aquí se continúa con la notación **w**.

42 Para simplificar la notación, p^* se escribe dentro del argumento de la divergencia KL en vez de $p(\mathbf{x} \mid \mathbf{D})$. Es una notación simplificada para la distribución a posteriori empleada en varios textos sobre el tema, y cuando aparezca en este escrito debe tomarse como tal.

mismos. Entonces se tiene la forma no normalizada de la distribución a posteriori que suele emplearse, como se mencionó en el capítulo referente al método bayesiano.

Considerando los cambios realizados a la función de optimización, la expresión puede ser presentada de varias formas, una muy conveniente es la siguiente:

$$\mathrm{KL}(q(\mathbf{w})\|p_1(\mathbf{w})) = \mathrm{KL}(q(\mathbf{w})\|p(\mathbf{D}|\mathbf{w})p(\mathbf{w})) = \mathrm{KL}(q(\mathbf{w})\|p(\mathbf{w})) + \mathrm{NLL}$$

Donde NLL es la **verosimilitud logarítmica negativa (*negative log-likelihood*),** es decir "- *log*(*p*(**D** | **w**))" que suele ser la función a minimizar durante el entrenamiento en una gran variedad de modelos de aprendizaje computacional. Con esta función se puede ver que en un contexto bayesiano de inferencia variacional, el modelo es entrenado conforme a la métrica tradicional que dicta cuán factibles son los datos considerando los parámetros usados (la NLL) y conforme a la métrica que dicta cuán similar es la distribución seleccionada a la distribución a priori (la divergencia KL entre $q(\mathbf{w})$ y $p(\mathbf{w})$).

De hecho, esta forma es la que se emplea en el código para los modelos desarrollados en los próximos capítulos. Aquí es importante no confundir KL($q(\mathbf{w})$ || $p(\mathbf{w})$) con KL(q || p^*), que es la divergencia con respecto a la distribución a posteriori normalizada p^*, mencionada previamente. También hay que notar que aunque $p(\mathbf{w})$ es la distribución a priori que se fija arbitrariamente, esto no quiere decir que siempre sea la misma durante la optimización, hay que recordar la capacidad de actualización del conocimiento del método bayesiano: la distribución a priori actual es la distribución a posteriori que se obtuvo tras considerar los últimos datos empleados en la inferencia.

En la documentación de TensorFlow Probability[43] para ciertos modelos, se indica que la función a minimizar se debe escribir en el código como KL($q(\mathbf{w})$ || $p(\mathbf{w})$) + NLL. Es decir, se obtienen ambos términos independientemente en el programa, y la función se define como su suma. En la estructura de los programas, esta forma es más intuitiva que obtener directamente KL($q(\mathbf{w})$ || $p_1(\mathbf{w})$). La documentación para los elementos usados en los programas no especifica cómo se aproxima el valor de KL($q(\mathbf{w})$ || $p(\mathbf{w})$) internamente. Es una característica integrada de los modelos probabilísticos de la plataforma, y simplemente debe "llamarse" para conocer su valor. El cálculo de la NLL se da de la misma forma, con funciones predefinidas, y no es necesario para el usuario derivar la expresión (*Tfp.Layers.DenseFlipout | TensorFlow Probability*, s. f.).

Inferencia de Monte Carlo:

Una **aproximación de Monte Carlo** consiste en generar muestras de alguna distribución, y aproximar dicha distribución con una distribución empírica (obtenida a partir de las muestras). De esta forma se pueden obtener los datos que se requieran con la distribución empírica, aunque la precisión siempre está sujeta al número de muestras generadas. Esta aproximación no está limitada a la "inferencia de Monte Carlo" e incluso se usa en algunos puntos de capítulos siguientes, que están basados en modelos de inferencia variacional (Murphy, 2012, pp. 52-56).

La inferencia de Monte Carlo consiste en obtener una aproximación de Monte Carlo de la distribución a posteriori, con una precisión que depende de las muestras generadas (mientras más muestras, más precisión). Es un método muy atractivo, pues en la inferencia variacional el procedimiento puede ser complicado en algunos casos, y además, a pesar de que sea un método mucho más robusto que la aproximación de Laplace, la precisión aún depende del tipo de distribución que se elija para la aproximación (Murphy, 2012, p. 815).

43 Es una de las herramientas principales usadas en este trabajo, y es un módulo de la plataforma TensorFlow. Tanto TensorFlow como TensorFlow Probability, y el rol de ambos en el desarrollo de este escrito, se describen con más detalle en el siguiente capítulo.

La inferencia de Monte Carlo resulta no ser tan simple como suena. Después de todo, ¿cómo se generan muestras de una distribución a posteriori cuya forma exacta no puede ser obtenida? Hay dos maneras para realizar este tipo de inferencia: una es con métodos no iterativos que generan muestras independientes, y otra es un método iterativo conocido como **Monte Carlo vía cadenas de Markov (*Markov Chain Monte Carlo* o MCMC)**, el cual, aunque genera muestras que no son independientes, tiene entre sus ventajas un buen rendimiento en espacios de muchas dimensiones. Los métodos varían en complejidad, recursos necesarios, etcétera; dependiendo de lo que se tenga al alcance (por ejemplo, en ocasiones se conoce la **función de densidad acumulativa (CDF)** de la distribución, y es más sencillo obtener muestras) o simplemente de lo que se requiera en el contexto de aplicación (Murphy, 2012, p. 815).

Mucho más podría decirse de los métodos de aproximación mencionados, pero ahondar en el tema sería poco práctico dado el enfoque de este trabajo. Se le dio un tratamiento un poco más profundo a la inferencia variacional porque es el método empleado en el trabajo, pero por lo general se dará mayor énfasis a los resultados obtenidos de los modelos bayesianos que a los conceptos detrás de su implementación computacional.

Una vez que se elige uno de los métodos descritos, se hace la aproximación, y el resultado es una distribución sobre los parámetros del modelo de regresión logística. Es decir, en vez de tener un valor estimado para cada uno de ellos, se tendrá una distribución que indicará un rango probabilístico estimado para los mismos; esto implica expresar la incertidumbre que se tiene con respecto a la estimación realizada. En el capítulo 7 se presenta un modelo de regresión logística empleando datos sintéticos, para ilustrar con un ejemplo práctico el método bayesiano y sus diferencias con el método tradicional.

REFERENCIAS

Choosing Logisitic Regression's Cutoff Value for Unbalanced Dataset. (s. f.). Recuperado 25 de mayo de 2020, de http://ethen8181.github.io/machine-learning/unbalanced/unbalanced.html

MacKay, D. J. C. (2003). *Information Theory, Inference, and Learning Algorithms*. Copyright Cambridge University Press.

MLPR w3a—Machine Learning and Pattern Recognition. (s. f.). Recuperado 24 de mayo de 2020, de https://www.inf.ed.ac.uk/teaching/courses/mlpr/2018/notes/w3a_intro_classification.html

MLPR w3c—Machine Learning and Pattern Recognition. (s. f.). Recuperado 25 de mayo de 2020, de https://www.inf.ed.ac.uk/teaching/courses/mlpr/2018/notes/w3c_logistic_regression.html

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT press.

Tfp.layers.DenseFlipout | TensorFlow Probability. (s. f.). TensorFlow. Recuperado 25 de mayo de 2020, de https://www.tensorflow.org/probability/api_docs/python/tfp/layers/DenseFlipout

Theodoridis, S. (2015). *Machine Learning: A Bayesian and Optimization Perspective* (1st ed.). Academic Press, Inc.

6. HERRAMIENTAS EMPLEADAS

En este trabajo, los modelos y experimentos se implementaron en el lenguaje de programación Python, usando también módulos, plataformas y entornos para este lenguaje. Antes de entrar en la descripción y el desarrollo de modelos, se describirán las herramientas usadas, con una justificación de su elección.

Python y Spyder: Python fue el lenguaje de programación elegido para este trabajo por dos razones principales, la cantidad de recursos disponibles relacionados con el tema y sus propias cualidades comparado a otros lenguajes. Por su parte, Spyder es el entorno de desarrollo elegido. Este entorno está orientado a científicos, ingenieros y analistas de datos, y es adecuado para los programas implementados y el trabajo en general. Spyder tiene varias ventajas y desventajas, pero en realidad la principal razón en la decisión fue la facilidad y rapidez de los procesos de instalación y configuración, que permiten implementar rápidamente los programas, con una interfaz intuitiva, e integrar los módulos necesarios con facilidad (*Spyder Website*, s/f).

Los recursos disponibles en Python son notables. Construir algoritmos de aprendizaje computacional con módulos básicos puede ser ineficiente y tedioso, cuando el único interés es implementar aplicaciones como modelos de clasificación. En estos casos es mejor emplear módulos de alto nivel que ya fueron optimizados a ese nivel básico. Considerando esto, para este trabajo se eligió Python en gran parte por ser la base de la plataforma TensorFlow (detalles más adelante), aunque también cabe destacar la presencia de otros módulos en el lenguaje como Keras, Pandas, NumPy, y Matplotlib.[44]

Python se eligió también por sus recursos en términos de la comunidad de usuarios: la popularidad de Python en el aprendizaje computacional es dominante. En términos concretos, esto se ha establecido investigando cuáles son los lenguajes más mencionados en ofertas de empleo en este campo (Puget, 2017), o los más usados en los proyectos de aprendizaje computacional disponibles en GitHub (“The State of the Octoverse”, 2019), entre otras estadísticas similares: el dominio de Python siempre es evidente. Gracias a esto hay una gran cantidad de recursos e implementaciones en Python, y una amplia comunidad que sigue creciendo.

En cuanto a las cualidades de Python como lenguaje, las ventajas más notables son: es simple comparado con otros lenguajes, no requiere compilación ni enlace (es un lenguaje interpretado), y su intérprete es interactivo (haciéndolo muy útil para experimentar). Como lo dice en la documentación de Python: los programas suelen ser mucho más cortos que sus equivalentes en C/C++ o Java, gracias a los tipos de datos de alto nivel, al agrupamiento mediante indentación en vez de llaves, y a que no necesita declaración de variables ni de argumentos (*1. Whetting Your Appetite — Python 3.8.3 documentation*, s/f).

Una alternativa que vale la pena mencionar es R. Cuando hay modelos probabilísticos y métodos bayesianos de por medio en una implementación, el lenguaje R (por lo general con el entorno R Studio) suele ser de las primeras opciones. Al igual que Python, R es un software libre, pero se centra en computación estadística, mientras que Python es más general en sus aplicaciones. Por lo general, Python es preferible si se considera la posibilidad de una etapa de producción para el proyecto, mientras que R suele emplearse si la prioridad es el análisis estadístico. Con esto en cuenta, R parecería una opción más natural, dado el contexto de investigación del presente trabajo, y además ofrece más facilidades en la presentación de resultados (principalmente en cuanto a representaciones gráficas). Sin embargo, Python suele tener código más fácil de leer y de mantener, además de ser más claro en la construcción de los algoritmos. Por el contrario R tiende más a

44 Keras es una API de aprendizaje profundo para Python, que aunque se puede usar independientemente, es parte de la estructura de TensorFlow (*Keras: the Python deep learning API*, s/f). Pandas es una librería de análisis de datos para Python (*pandas - Python Data Analysis Library*, s/f). NumPy es un paquete esencial para computación científica en Python (*NumPy*, s/f). Matplotlib es una librería para crear visualizaciones diversas en Python (*Matplotlib: Python plotting — Matplotlib 3.2.1 documentation*, s/f).

"ocultar" elementos detrás de funciones predefinidas (*R Vs Python: What's the Difference?*, s/f) (Kan, 2018), (*R: What is R?*, s/f).

Ya que este trabajo se centra en los resultados más que en el código, y por el moderado grado de complejidad de los programas elaborados, cualquiera de estas dos opciones hubiera sido adecuada, pero en esta ocasión se optó por Python, pues su popularidad y extenso crecimiento hacen más factible que la mayoría de las aplicaciones relacionadas al tema, en un futuro, estén basadas en Python. Algunos factores adicionales que favorecen a Python, y que están relacionados con su dominio en cuanto a popularidad son: módulos bien implementados y ampliamente documentados para aprendizaje profundo (*deep learning*)[45], más facilidades en cuanto a accesibilidad y reproducibilidad (esto implica que es mejor para presentar resultados en sitios web o aplicaciones), y un rendimiento superior: es un lenguaje considerablemente más rápido para varias tareas importantes en el aprendizaje computacional y la ciencia de datos, como la carga de datos y el *bootstrapping*[46] (Kan, 2018), *(R Vs Python: What's the Difference?, s/f)*.

Como conclusión, vale la pena mencionar la opinión derivada de un estudio que encuestó a miles de desarrolladores de aprendizaje computacional, donde, aunque se reconoció el dominio en popularidad de Python, también se hizo notar que entre aplicaciones específicas (detección de fraudes, procesamiento de lenguaje natural, seguridad de redes, etc.) no siempre es la norma (por ejemplo, C/C++ es una opción más popular para la inteligencia artificial en juegos). En ese estudio se concluye que la decisión se debe basar en la aplicación, la experiencia profesional del desarrollador mismo, y su razón para involucrarse con el aprendizaje computacional (Economics, 2019).

TensorFlow y TensorFlow Probability: se mencionó que Python fue elegido para este trabajo en gran parte por contener la herramienta TensorFlow. Del mismo modo, durante la primera etapa de la investigación, el módulo Edward tuvo mucho que ver con la elección de TensorFlow para este trabajo (a pesar de que los modelos finalmente se implementaron sin usar el módulo Edward).

Entre las herramientas de alto nivel para el aprendizaje computacional, dos de las plataformas más populares, TensorFlow y PyTorch, están disponibles principalmente en Python[47] (*API Documentation | TensorFlow Core v2.2.0*, s/f), (*PyTorch*, s/f). Tanto TensorFlow como PyTorch son plataformas dominantes en el tema, y elegir TensorFlow se basó más bien en la diferencia entre las herramientas Edward y Pyro. Edward funciona sobre TensorFlow, y Pyro sobre PyTorch: ambos son lenguajes de programación probabilística, que permiten escribir modelos bayesianos y obtener el resultado automáticamente para los datos de interés, sin la necesidad de derivar manualmente el algoritmo de inferencia (que suele ser una tarea demandante). Ambos módulos tienen herramientas más que suficientes para las tareas requeridas en este trabajo, pero se optó por TensorFlow porque actualmente el módulo TensorFlow Probability ha integrado Edward2 dentro de su estructura en "capas"[48], simplificando el uso del módulo en esta plataforma. Finalmente

45 En el aprendizaje computacional, los modelos empleados usan la representación de la información determinada durante el diseño, y dependen de ella para su funcionamiento. El aprendizaje profundo se funda sobre la idea de usar representaciones expresadas en términos de otras representaciones más sencillas, para eliminar la dependencia de una representación provista, y permitir que la máquina construya conceptos complejos por su cuenta (Goodfellow et al., 2016, pp. 1–8). En el presente trabajo no se lidia con el aprendizaje profundo, pero TensorFlow, la plataforma fundamental para este trabajo, es uno los principales módulos para aprendizaje profundo.

46 El *bootstrapping* es un método estadístico que toma muestras aleatorias de una población (Murphy, 2012, pp. 192–193).

47 Pueden emplearse con otros lenguajes, pero sería necesario un *binding* (adaptación al lenguaje deseado) (*Standards, APIs, Interfaces and Bindings*, 2015), sin mencionar que la gran mayoría de recursos y documentación están siempre enfocados al lenguaje Python.

48 Donde la base o primera capa es TensorFlow, la segunda capa se refiere a bloques de construcción estadísticos, la tercera a la construcción de modelos (aquí entra Edward2) y la última se refiere a la inferencia probabilística (*TensorFlow Probability*, s/f).

no fue necesario usar Edward explícitamente, pues las herramientas requeridas para este trabajo ya fueron integradas a TensorFlow Probability, aún sin hacer uso del módulo (*Edward – Home*, s/f), (*Module*, s/f), (*Pyro*, s/f).

Otra motivación para elegir TensorFlow y TensorFlow Probability, y no otras herramientas, fue que en esta tesis se busca darle cierto tratamiento a la estructura de los modelos, y a elementos que se deben considerar dentro de la inferencia. En este sentido, TensorFlow Probability es preferible a paquetes como JAGS en R o PyMC en el mismo Python, donde los modelos se especifican en un nivel aún más alto, cubriendo automáticamente varios detalles que se abordan en este trabajo, y de forma explícita en el código. De hecho, PyMC4 (la última versión de PyMC) se implementará usando TensorFlow Probability (Developers, 2019).

Hardware: todas las implementaciones fueron desarrolladas y ejecutadas en una computadora con procesador Intel Core-i7 (cuarta generación) 4810MQ de 2.8GHz, memoria RAM de 16GB con tecnología DDR3L SDRAM, y con sistema operativo Windows 8.1. Esta decisión es de poca importancia para los objetivos del trabajo, que en ningún momento se relacionan con optimizar los procesos implementados. En general, las tareas que se llevan a cabo no son lo suficientemente elaboradas para considerar las ventajas de emplear hardware dedicado. No es inusual que los modelos, especialmente en el aprendizaje profundo, sean de tal complejidad, que entrenarlos lleve hasta días, y requieran un poder computacional considerable, también debido a enormes cantidades de datos. Este no es el caso para los modelos implementados en este trabajo, que además de tener pocos parámetros, sólo tienen una cantidad moderada de datos a su disposición.

Aunque sí se consideran los efectos en el rendimiento al emplear un método bayesiano comparado con un esquema tradicional, se adquiere una perspectiva general, conceptual y metodológica, diferente al análisis de métricas de rendimiento computacional (como el tiempo de ejecución) para los modelos, implementaciones, y hardware específicos que se usan en este trabajo. Sería necesario un meticuloso análisis (más allá del enfoque de este escrito) para aportar conclusiones innovadoras y generales considerando esta segunda perspectiva.

No se empleó ningún hardware especializado porque no es requerido para el funcionamiento del software empleado, ni tendría un impacto significativo para las tareas desarrolladas.

REFERENCIAS

1. Whetting Your Appetite—Python 3.8.3 documentation. (s/f). Recuperado el 29 de mayo de 2020, de https://docs.python.org/3/tutorial/appetite.html

API Documentation | TensorFlow Core v2.2.0. (s/f). TensorFlow. Recuperado el 29 de mayo de 2020, de https://www.tensorflow.org/api_docs?hl=es

Developers, P. (2019, junio 1). *Theano, TensorFlow and the Future of PyMC*. Medium. https://medium.com/@pymc_devs/theano-tensorflow-and-the-future-of-pymc-6c9987bb19d5

Economics, D. (2019, diciembre 11). *What is the best programming language for Machine Learning?* Medium. https://towardsdatascience.com/what-is-the-best-programming-language-for-machine-learning-a745c156d6b7

Edward – Home. (s/f). Recuperado el 29 de mayo de 2020, de http://edwardlib.org/

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Kan, C. E. (2018, diciembre 10). *Data Science 101: Is Python better than R?* Medium. https://towardsdatascience.com/data-science-101-is-python-better-than-r-b8f258f57b0f

Keras: The Python deep learning API. (s/f). Recuperado el 15 de junio de 2020, de https://keras.io/

Matplotlib: Python plotting—Matplotlib 3.2.1 documentation. (s/f). Recuperado el 15 de junio de 2020, de https://matplotlib.org/3.2.1/index.html

Module: Tfp.edward2 | TensorFlow Probability. (s/f). TensorFlow. Recuperado el 29 de mayo de 2020, de https://www.tensorflow.org/probability/api_docs/python/tfp/edward2

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT press.

NumPy. (s/f). Recuperado el 15 de junio de 2020, de https://numpy.org/

pandas—Python Data Analysis Library. (s/f). Recuperado el 15 de junio de 2020, de https://pandas.pydata.org/

Puget, J.-F. (2017, marzo 27). *The Most Popular Language For Machine Learning Is …*. Medium. https://medium.com/inside-machine-learning/the-most-popular-language-for-machine-learning-is-46e2084e851b

Pyro. (s/f). Recuperado el 29 de mayo de 2020, de https://pyro.ai/

PyTorch. (s/f). Recuperado el 29 de mayo de 2020, de https://www.pytorch.org

R Vs Python: What's the Difference? (s/f). Recuperado el 15 de noviembre de 2019, de https://www.guru99.com/r-vs-python.html

R: What is R? (s/f). Recuperado el 29 de mayo de 2020, de https://www.r-project.org/about.html

Spyder Website. (s/f). Recuperado el 29 de mayo de 2020, de https://www.spyder-ide.org/

Standards, APIs, Interfaces and Bindings. (2015, enero 16). https://web.archive.org/web/20150116081559/http://www.acm.org/tsc/apis.html

TensorFlow Probability. (s/f). TensorFlow. Recuperado el 29 de mayo de 2020, de https://www.tensorflow.org/probability/overview?hl=vi

The State of the Octoverse: Machine learning. (2019, enero 24). *The GitHub Blog*. https://github.blog/2019-01-24-the-state-of-the-octoverse-machine-learning/

7. ASPECTOS PRÁCTICOS DE MODELOS DE APRENDIZAJE COMPUTACIONAL Y EL MÉTODO BAYESIANO: INFERENCIA SOBRE DATOS SINTÉTICOS

Antes de presentar el problema central de este trabajo, junto con las implementaciones requeridas, se tomará un ejemplo con datos sintéticos, basado en el código del ejemplo para regresión logística bayesiana elaborado por los autores de TensorFlow Probability (*Tensorflow/Probability*, s. f.). La idea es presentar varios conceptos y la esencia del método bayesiano en una forma práctica y fácil de visualizar. Así se agilizará la discusión y se puntualizará el enfoque cuando se desarrollen los modelos correspondientes al problema del prejuicio en el aprendizaje computacional.

El ejemplo consiste en generar de manera aleatoria los parámetros de una regla de clasificación, es decir, la frontera que especifica qué elementos corresponden a qué clase. Para facilitar la visualización, esta frontera se coloca en el espacio bidimensional, por lo que sólo hay dos parámetros. Con estos parámetros "verdaderos" se invierte la regla de clasificación, y añadiendo ruido a los resultados, se generan observaciones ruidosas. La tarea consiste en tratar de recuperar la regla verdadera de clasificación, a partir de las observaciones, usando regresión logística, que es el modelo empleado en el resto del trabajo.

En este ejemplo se revisarán las decisiones básicas para el modelo una vez que ya se eligió su forma general (regresión logística en este caso). Estas decisiones no dependen de si el modelo es bayesiano o no, pues todos los modelos tienen el fin de ser "entrenados" con datos, para ajustar sus parámetros y modelar el proceso que los generó con la mayor precisión posible (sin perder de vista la generalización[49]).

Las Fases del Modelo Clasificador (o cualquier modelo)

Hay varias maneras de expresar los pasos a seguir al elaborar modelos de aprendizaje computacional y modelos estadísticos. La cantidad de fases definidas varían acorde a la aplicación considerada. En este caso, con el apoyo de varias fuentes, se expresarán sólo dos fases cruciales para el desarrollo de este trabajo: preprocesamiento de datos y modelado. En cada uno de estos pasos, se integrarán fases que pueden considerarse pasos independientes. Esto se hace con el afán de no desviar el enfoque del trabajo hacia las especificaciones de los modelos, pues podrían desarrollarse investigaciones extensas acerca de cada una de ellas. A continuación se esboza cómo se explorará cada fase y los puntos relevantes para este trabajo, tomando como base las siguientes referencias: (Shiflet & Shiflet, 2014, pp. 7-11), (Posted by Eddie Soong on June 29 & Blog, s. f.), **(***Modeling process - Universidad de California en Santa Cruz*, s. f.**).**

Preprocesamiento de Datos

En este punto, los datos a usarse son explorados para idear un modelo adecuado. Antes también debe planearse adecuadamente la recolección de datos, pero es una etapa que no se revisa en este trabajo: los datos empleados se consultan directamente o bien, se generan, como es el caso en este ejemplo. En cuanto a la exploración, para estos datos sintéticos no tiene sentido, pues ya se conocen los detalles de cómo fueron generados; en capítulos siguientes se podrá apreciar que hay información valiosa que puede obtenerse durante esta parte del proceso.

Aunque en este trabajo el enfoque está sobre el modelado y no sobre la recolección y minería de datos, es un área que también tiene una fuerte relación con la temática principal. Los métodos bayesianos tienen como una más de sus ventajas la capacidad de lidiar con datos

49 El término *overfitting* se usa en el contexto del aprendizaje computacional para hacer referencia a un ajuste excesivamente preciso del modelo para los datos de entrenamiento. Es decir, los parámetros del modelo explican a la perfección todos los datos que éste tiene al alcance, pero por ese ajuste tan rígido es probable que haga malas predicciones para datos nuevos en puntos que no ha visto (o incluso que ya vio, pues al confiar ciegamente en los datos de entrenamiento, si alguno de ellos fuera una observación mal capturada o ruidosa, las predicciones para futuras observaciones con entradas similares serían imprecisas) (Murphy, 2012, p. 22).

incompletos mediante principios estadísticos (Ma & Chen, 2018). Del mismo modo, los modelos generativos tienen una importante ventaja sobre los discriminativos en situaciones con datos incompletos (Murphy, 2012, p. 268) (en el capítulo sobre clasificadores se explicó la diferencia entre estos dos tipos de modelo). Incluso en la aplicación central de esta tesis (lidiar con el sesgo dentro del aprendizaje computacional), se han elaborado trabajos importantes enfocados completamente a la etapa de preprocesamiento de datos[50].

Modelado

Una vez que se exploran los datos, es necesario postular el modelo a emplear. No se hace énfasis en este paso, pues como se verá más adelante, el desarrollo práctico se hace sobre un caso de estudio que usa modelos de regresión logística, por lo que la decisión ya fue tomada en lo que concierne a este escrito. Además, el objetivo central es observar las diferencias al considerar un esquema bayesiano: al usar el modelo de regresión logística en la totalidad de los experimentos, es sencillo hacer énfasis en estas diferencias.

Con un modelo ya seleccionado, viene la fase de ajustar el modelo a los datos. Aunque tampoco es el enfoque del trabajo, se revisará el tema, pues es una fase indispensable de todos los programas elaborados en este escrito. Esta fase implica seleccionar una función de costo a minimizar, así como un proceso de optimización con sus correspondientes cualidades. Ya se habló de algunas funciones de costo en el capítulo referente a clasificadores, como la divergencia KL o la verosimilitud logarítmica negativa (NLL). Por otra parte, hay una variedad de optimizadores para minimizar la función de costo (independientemente de si el modelo es bayesiano o no), y tampoco se le prestará mucha atención a este tema. Aunque se discutirán y justificarán las decisiones tomadas, observar el efecto de usar distintos optimizadores o funciones de costo en la precisión o en el sesgo final de los modelos no es el objetivo en este trabajo.

Tras entrenar un modelo, se puede evaluar el mismo, y si no se obtienen resultados adecuados, es hora de regresar al primer paso del modelado, la postulación del modelo. El proceso es iterativo hasta obtener los resultados deseados, y una vez que se obtienen, el modelo se implementa para cualquiera que sea la aplicación. En este trabajo el enfoque está en este último punto del modelado, la evaluación de los modelos: se compararán las respuestas que ofrecen las versiones clásica y bayesiana del modelo.

Típicamente, el punto de comparación entre modelos es su rendimiento al considerar nuevos datos que no estuvieran disponibles durante las fases de entrenamiento y formulación del modelo (Barber, 2012, pp. 306-307). En este caso, dicha métrica no es el único punto de comparación, pues más allá de qué tan precisas son las predicciones de los modelos numéricamente, el interés está sobre la forma en que cada modelo ofrece sus respuestas: ¿el modelo da una simple predicción numérica u ofrece también una cuantificación o medida de la incertidumbre sobre sus respuestas?, ¿cómo se pueden interpretar estadísticamente las respuestas del modelo?, ¿cómo son relevantes las preguntas anteriores en aplicaciones que involucran sesgo o prejuicio, y discriminación?

50 Un ejemplo detallado de este enfoque es el trabajo de (Calmon et al., 2017).

7.1 DESARROLLO

Tarea de Clasificación y Datos Sintéticos

Como ya se mencionó antes, la presente tarea consiste en hallar la regla de clasificación para los datos sintéticos generados. Estos datos son bidimensionales, por lo que están descritos únicamente por dos parámetros: x_1 y x_2. Teniendo control sobre los datos generados, se puede escoger arbitrariamente la cantidad de datos disponibles para entrenar el modelo, y más adelante se observará cuál es el efecto de tener disponibles más o menos datos. Estableciendo una semilla aleatoria o *seed*, se asegura que el programa de generación dará datos reproducibles para las diferentes iteraciones del programa. A continuación se tienen varias cantidades de datos generados, ilustrando también la "frontera verdadera" que se empleó para generar los puntos en el plano: nótese que no todos los puntos están del lado correcto de la frontera, y ello se debe a la adición de ruido para tener una tarea de clasificación realista.

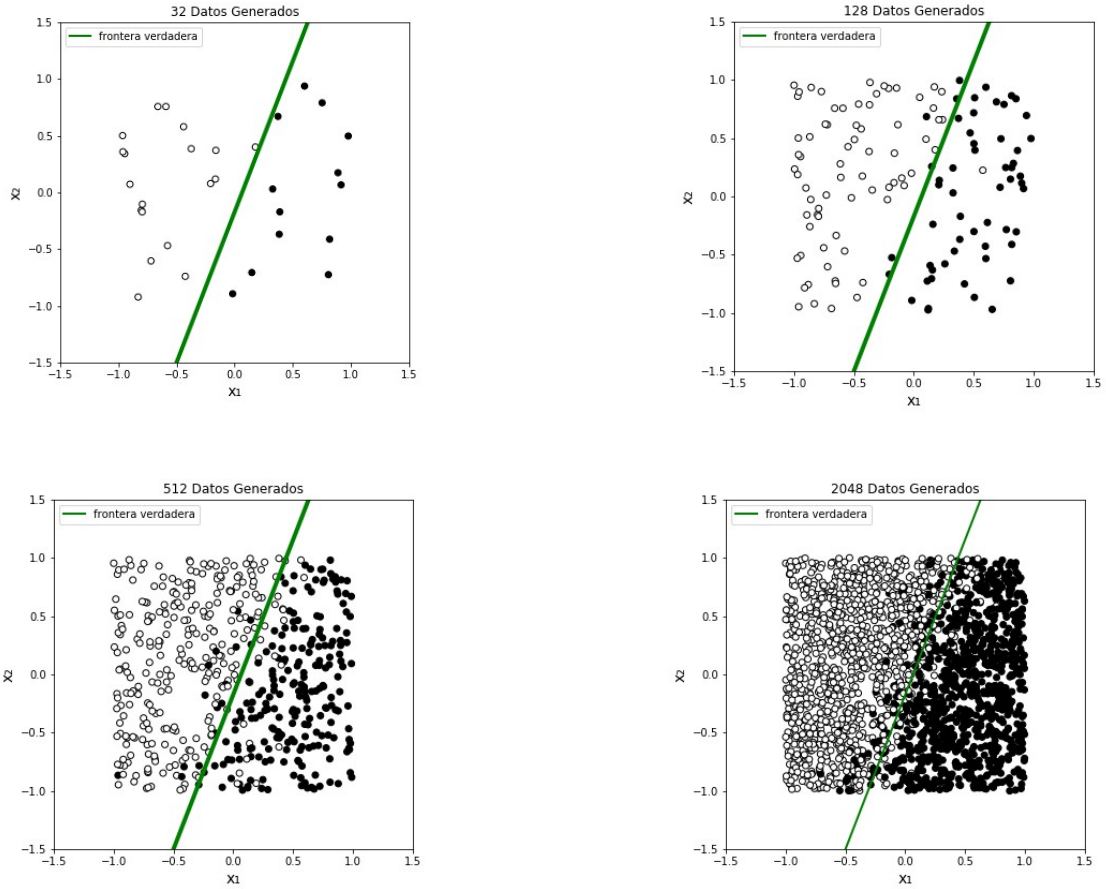


Figura 7.1. Generación de datos sintéticos: se puede apreciar la frontera verdadera a inferir con los modelos de regresión logística, y también los datos que se ocuparán para la tarea. Conforme hay más datos, es más evidente que varios de ellos no están del lado correcto de la frontera, esto es gracias al ruido añadido (la regla original sin ruido, colocaría a todos los puntos negros del lado derecho de la frontera y a los blancos del lado izquierdo). Al tener más datos, las clases o grupos se aprecian con mayor claridad.

En un caso real, únicamente se dispone de los puntos o datos, y el objetivo es inferir la frontera verdadera. Ese es el objetivo en este ejemplo: a continuación se emplearán modelos de regresión logística en distintos escenarios, para observar cómo cada uno de ellos estima la regla de clasificación (la frontera entre puntos blancos y puntos negros). Ello implica deducir cómo las covariables **X**, mediante los parámetros **w**, determinan la clase de cada punto.

La frontera representa la región donde la probabilidad de pertenecer a las clases es igual. Al observar nuevos puntos, los que estén sobre la línea podrían pertenecer a cualquier clase con probabilidades idénticas.

Regresión Logística

Ahora se usa el modelo de regresión logística para estimar la frontera verdadera conociendo únicamente los datos. Ya que en la regresión logística (bayesiana o no) la clasificación se basa en la probabilidad de un modelo de Bernoulli, la **verosimilitud** de los parámetros es la probabilidad $p(\mathbf{y} \mid \mathbf{X}, \mathbf{w})$ bajo dicho modelo, donde la probabilidad de éxito está modelada mediante la función sigmoide de la combinación lineal de covariables **X**, especificada por los parámetros **w** (y desplazada de acuerdo al término constante *b*). Los parámetros se optimizan al maximizar esta verosimilitud, o bien, al minimizar la **NLL** (verosimilitud logarítmica negativa), que es la función de costo a usar.

A continuación se muestran las aproximaciones a la frontera verdadera, obtenidas mediante la minimización de la NLL, para distintas cantidades de datos disponibles:

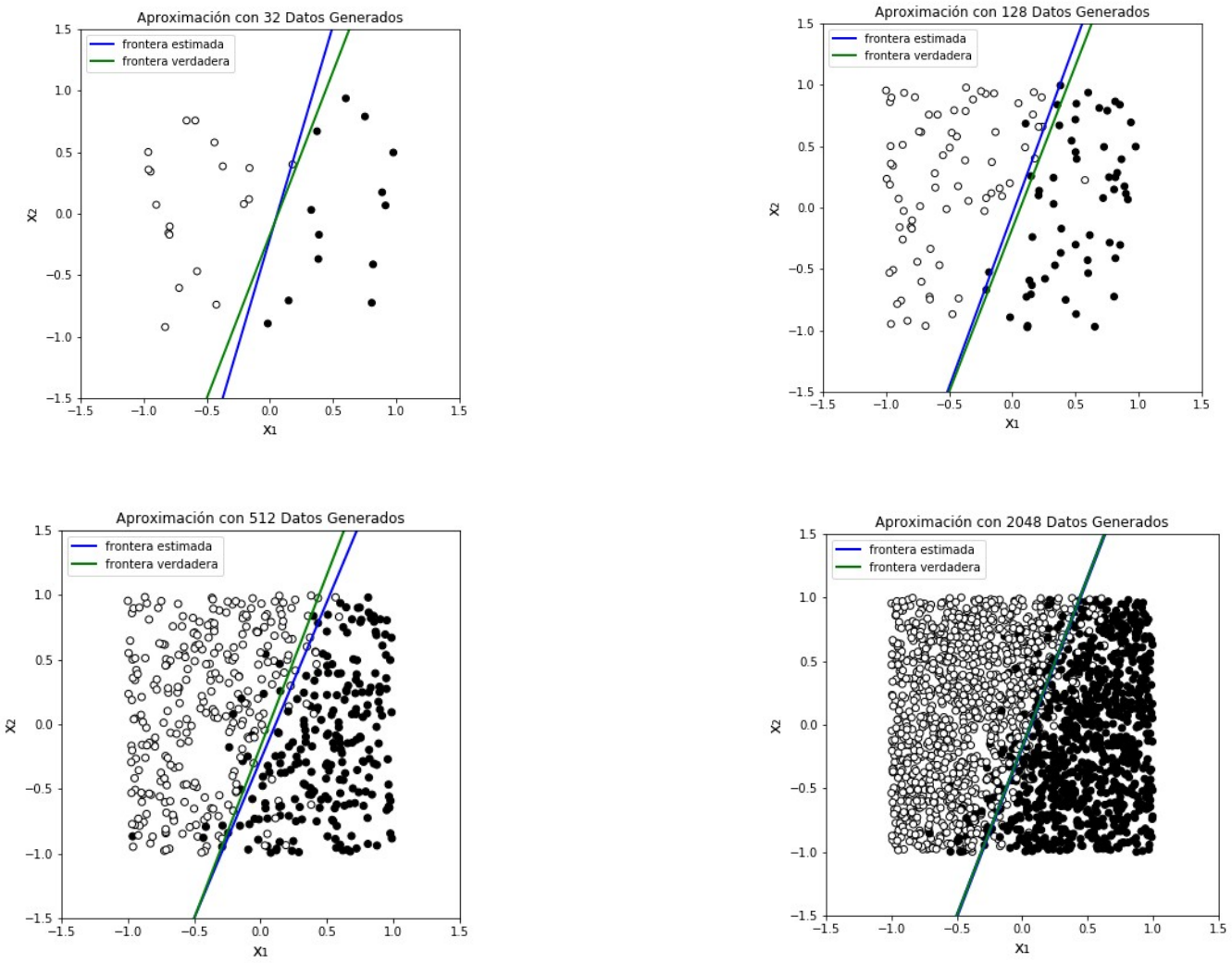


Figura 7.2. Fronteras estimadas para distintas cantidades de datos.

Como se puede observar, la tendencia es que se tienen aproximaciones más precisas conforme aumenta la cantidad de datos disponibles. Esto no es una norma absoluta, pues como se puede observar en el caso de 512 datos, la cercanía entre la frontera estimada y la verdadera es muy parecida a la que se observa para el caso con 128 datos. En la práctica no se conoce la frontera verdadera, así que sería natural asumir que la aproximación con 512 datos es más precisa, sin embargo hay heurísticas que se basan en frenar el proceso de optimización para evitar que el modelo se "deje llevar" por los datos, en casos donde el modelo generaliza mejor si no termina el entrenamiento (el método se conoce como *early stopping*) (Murphy, 2012, p. 263).

Regresión Logística Bayesiana

Ahora se planteará el problema con un esquema bayesiano de inferencia variacional. Como se mencionó en el capítulo de clasificadores, en el caso bayesiano la función de costo es nuevamente la NLL, pero se le añade también la divergencia KL entre la distribución a priori y la distribución variacional[51]. Como la solución bayesiana ya no es una simple frontera estimada, sino una distribución de ellas, se visualizarán varias muestras, donde, de acuerdo a la metodología bayesiana, cualquiera de ellas podría ser la frontera verdadera (con distintas probabilidades). La solución bayesiana es la distribución a posteriori, pero también es necesario definir la distribución a priori, es decir, la distribución de las fronteras que se asumen como posibles antes de observar cualquier dato; se puede decir que esta distribución es la frontera estimada para 0 datos generados:

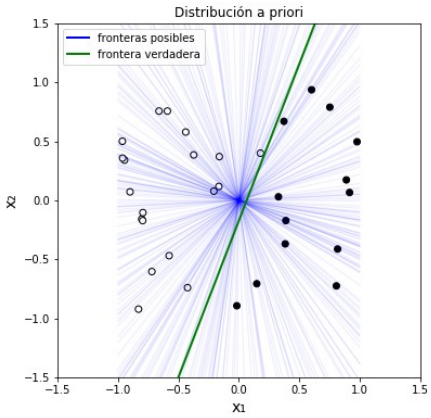


Figura 7.3. Distribución a priori de fronteras (parámetros) posibles.

En la Figura 7.3 se puede apreciar el significado de la distribución a priori: antes de considerar los datos, el conocimiento actual del modelo expresa solamente la suposición (o conocimiento previo) de que la frontera será una línea en el plano de las covariables **x**, y del mismo modo se puede observar que se asume que la frontera pasará por el origen o cerca del mismo. Aunque en la Figura 7.3 se muestran los primeros 32 datos, esto sólo es ilustrativo, pues las fronteras posibles mostradas las produce el modelo antes de ver los datos.

A continuación se muestran las distribuciones a posteriori, con sus medias, para distintas cantidades de datos observados:

51 Recordando ese capítulo, la función de costo es la divergencia KL entre la distribución a posteriori no normalizada y la distribución variacional, pero se puede expresar también como la suma mencionada aquí.

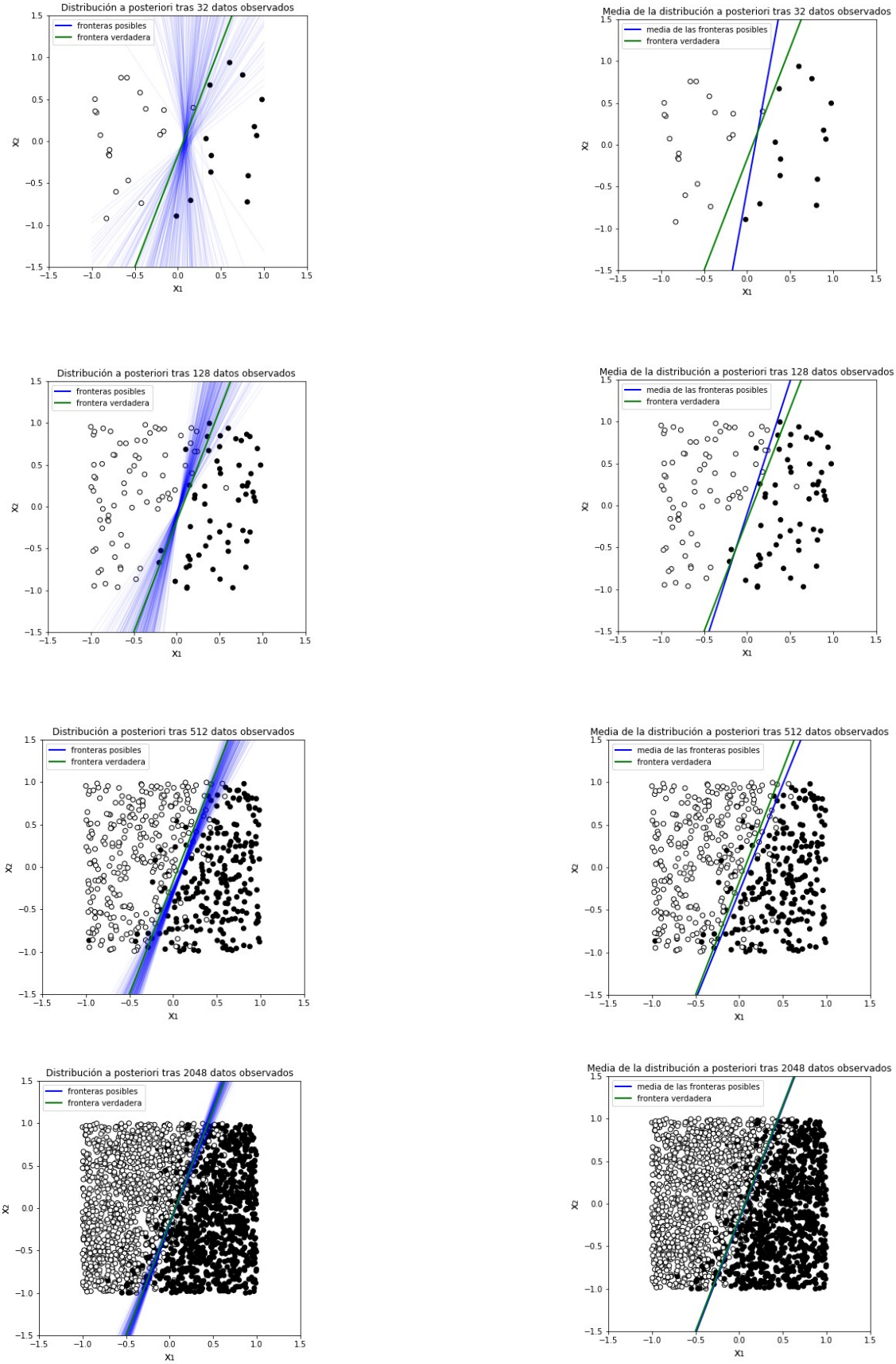


Figura 7.4.1. Distribuciones a posteriori y sus medias para distintas cantidades de datos observados.

Así como la frontera estimada en el caso no bayesiano se va aproximando a la frontera verdadera conforme se tienen más datos, el mismo efecto se puede ver en la media de las distribuciones a posteriori del caso bayesiano. Aunque hay toda una distribución de fronteras posibles, tiene lógica emplear la media si es necesario estimar la frontera con un solo valor puntual, pues es el punto donde se concentra la mayor densidad de probabilidad de la distribución.

La media a posteriori parece tener una precisión similar a la frontera estimada del caso no bayesiano, aunque para el caso con 32 datos se puede observar una mayor cercanía de la estimación no bayesiana a la frontera verdadera. A pesar de esto, considerando la distribución a posteriori y no sólo su media, es notable la ventaja del método bayesiano: conforme se observan más datos, no sólo se aproxima más a la frontera verdadera, también se reduce la dispersión de las fronteras posibles, es decir, la varianza de la distribución a posteriori. Esto es una medida del rango en que se considera que podría estar la frontera, así que se tiene una noción concisa de la incertidumbre.

En un escenario donde fuera necesario clasificar nuevos datos mediante la frontera estimada, el método bayesiano puede ofrecer una medida de cuánta certeza hay sobre el estimado del modelo, lo cual no está presente en el caso frecuentista. Por ejemplo, considerando 32 datos, la frontera estimada por el método no bayesiano parece estar más cerca de la frontera verdadera que la media del método bayesiano, por lo que si hay interés en predecir la clase de un nuevo valor, la frontera estimada del primer método será más precisa. Por su parte, el método bayesiano haría predicciones basadas en la media a posteriori, ligeramente más imprecisas, pero que también pueden hacer explícita la gran incertidumbre que hay en varios puntos del plano para esta cantidad de datos.[52]

Otra gran ventaja es que la medida de la incertidumbre no sólo describe cuán seguro está el modelo de sus respuestas considerando la información actual, sino que también depende de las características de dicha información: es decir, hay puntos en el plano donde la incertidumbre es mayor que en otros. En las distribuciones a posteriori de la Figura 7.4.1, se puede notar que por lo general no hay mucha incertidumbre cerca del origen, pero hay bastante hacia los extremos de la frontera: esto es porque siempre habrá más incertidumbre en las áreas en que se hayan visto menos datos. Es decir, si nunca se ha visto la clase de un punto (x_1, x_2), ni de puntos cercanos a éste, el modelo no tendrá certeza al dar una predicción para ese punto. Esta es una característica muy ventajosa, pues en el contexto de analizar datos cuya recolección es complicada o costosa, el método bayesiano ofrece cierta "ubicación" de la incertidumbre, y puede entonces usarse para determinar en qué zonas del espacio de covariables vale la pena recolectar más datos o realizar más experimentos, y así disminuir la incertidumbre de una forma eficaz (*MLPR w7b - Machine Learning and Pattern Recognition*, s. f.).

Este ejemplo permite una visualización de las ventajas del método bayesiano. En el caso de estudio central de los próximos capítulos, se estudiará cómo estas características se transfieren a la aplicación del método dentro del sesgo y la justicia en modelos de aprendizaje computacional (el modelo seguirá siendo el de regresión logística). En ese contexto habrá más covariables en juego que en este caso sintético, a tal grado que ya no será posible la visualización explícita de la frontera.

52 Como puede observarse para mayores cantidades de datos, no siempre hay mayor precisión estimando con el modelo no bayesiano que haciéndolo con la media del método bayesiano. En este ejemplo particular, para 32 datos, así sucedió, y se mencionó sólo para comentar que a pesar de que el método bayesiano no siempre será superior en la precisión de la predicción puntual, siempre incorporará la incertidumbre en sus resultados de una forma que el método frecuentista no puede.

Detalles Adicionales de los Modelos

Algunas decisiones al entrenar los modelos se han omitido, pues quedan fuera del enfoque de este trabajo, sin embargo vale la pena mencionarlas (y justificarlas):

Entrenamiento por mini-lotes:

La optimización de modelos de aprendizaje computacional se lleva a cabo mediante descenso de gradiente o método del gradiente. Esto quiere decir que la función de costo se minimiza obteniendo su derivada, para seleccionar parámetros en la dirección en que la función de costo decrezca con mayor rapidez o razón de cambio. Existen tres formas de llevar a cabo el proceso: por lote, donde se obtiene la dirección de cambio adecuada empleando todos los datos disponibles en el modelo, y después se actualizan los parámetros, reiterando hasta la convergencia; de forma estocástica, donde para cada dato disponible, se obtiene la dirección y se actualizan los parámetros, antes de pasar al siguiente dato; y por mini-lotes, donde se define un tamaño fijo de mini-lotes, y entonces se obtiene la dirección de cambio para todo un mini-lote y se actualizan los parámetros antes de pasar al siguiente mini-lote.

El método estocástico es el más robusto de los tres, y usualmente tendrá una mejor generalización, sin embargo, es computacionalmente costoso, pues requiere actualizar el modelo después de cada dato. El método por lote es el más eficiente, pues es mucho menos costoso actualizar parámetros una sola vez para todos los datos en cada iteración. El método por mini-lotes es un balance entre la eficiencia del segundo y la robustez del primero: si el tamaño de mini-lote equivale a todos los datos, simplemente es el método por lote, y si equivale a uno, entonces es el método estocástico. A través de este trabajo se emplea un método por mini-lotes con mini-lotes de 32 datos (una recomendación práctica es siempre usar mini-lotes de 32 o menos datos) (Masters & Luschi, 2018).

Funciones de costo:

Dependiendo del modelo se tendrán diferentes funciones de costo. Como ya se mencionó, la verosimilitud logarítmica negativa es una elección natural para los modelos de aprendizaje computacional (incluyendo la regresión logística), ya que representa matemáticamente cuán probables son los datos que se han observado para los parámetros del modelo. Y a su vez, la divergencia KL se emplea en la regresión logística bayesiana (y de forma general en modelos de inferencia variacional), para determinar una distancia entre la distribución verdadera y la distribución variacional.

Podrían emplearse otras funciones de costo, como la diferencia cuadrada entre salidas calculadas con el modelo para las entradas observadas f(**x**, **W**), y los valores de salida calculados **y** (esta es una función de costo muy común en regresión lineal). También podrían añadirse regularizadores, entre otros elementos, a la función de costo, que representan matemáticamente cualidades deseadas para el modelo.

Por ejemplo, en este caso sintético, dentro del proceso de optimización ambos tipos de errores son penalizados de la misma forma, es decir, clasificar erróneamente a un elemento nuevo dentro de la clase de los puntos blancos no es más ni menos grave que cometer el mismo error para la clase de puntos negros. En ciertas aplicaciones, sí habrán diferencias entre errores, y éstas se pueden integrar en la función de costo para que el modelo procure evitar cierto tipo de errores más que otros[53]. De forma similar, los regularizadores son términos que se añaden a la función de costo para reducir la magnitud o el número efectivo de parámetros. El regularizador L2 es muy común dentro de la regresión lineal, y consiste en añadir un término a la función de costo: la magnitud del vector de parámetros multiplicada por un escalar, dicho escalar controla qué tanto el modelo priorizará mantener valores de baja magnitud para sus parámetros (Murphy, 2012, pp. 225-227).

53 Esto ya se discutió en el capítulo correspondiente a clasificadores, al discutir el concepto de riesgo promedio.

Todas estas posibles variantes para las funciones de costo, no son irrelevantes, y serán contempladas en el desarrollo. Sin embargo, no se explorarán a fondo en las implementaciones siguientes, pues la comparación entre un modelo bayesiano y otro que no lo es, se simplifica al emplear en ambos casos funciones de costo comunes.

Optimizador Adam:

Con una función de costo definida y un esquema de entrenamiento seleccionado (en este caso por mini-lotes) *sólo* queda minimizarla. Hay una variedad de opciones para implementar la minimización de una función de costo, que son efectivas y eficientes en distintos contextos, y son el resultado del extenso trabajo previo en el tema.

Cada optimizador será más adecuado para algunas aplicaciones específicas, dependiendo de varios factores, como la tasa de aprendizaje[54] y su configuración, o el esparcimiento de los datos. El optimizador de estimación adaptada de momento (a*daptive moment estimation* o Adam) es el empleado a lo largo del trabajo. Esta decisión está basada en que Adam tiene un rendimiento promedio bastante bueno en comparación a las demás opciones, además es una opción bastante común y está integrado dentro de TensorFlow (Ruder, 2016).

Seleccionar un optimizador para cierto tipo de modelo o datos y hacer comparaciones entre optimizadores es un proceso que puede ser exhaustivo y desviar en gran manera el enfoque de este trabajo (la prioridad es comparar el modelo bayesiano con el que no lo es, y no el rendimiento general de los modelos para distintos optimizadores). Es por ello que se opta por Adam como la decisión fija en las implementaciones anteriores y próximas.

REFERENCIAS

Barber, D. (2012). *Bayesian reasoning and machine learning*. Cambridge University Press.

Calmon, F. P., Wei, D., Ramamurthy, K. N., & Varshney, K. R. (2017). *Optimized Data Pre-Processing for Discrimination Prevention*.

Ma, Z., & Chen, G. (2018). Bayesian methods for dealing with missing data problems. *Journal of the Korean Statistical Society*, *47*. https://doi.org/10.1016/j.jkss.2018.03.002

Masters, D., & Luschi, C. (2018). *Revisiting Small Batch Training for Deep Neural Networks*.

MLPR w7b—Machine Learning and Pattern Recognition. (s. f.). Recuperado 31 de mayo de 2020, de https://www.inf.ed.ac.uk/teaching/courses/mlpr/2018/notes/w7b_gaussian_processes.html

Modeling process—Universidad de California en Santa Cruz. (s. f.). Coursera. Recuperado 1 de octubre de 2019, de https://www.coursera.org/learn/mcmc-bayesian-statistics/lecture/vBMAX/modeling-process

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT press.

54 La tasa de aprendizaje es un hiper-parámetro, que determina la magnitud de "los pasos" o distancias que el optimizador contempla para "mover" los parámetros en cierta dirección, mientras busca los valores óptimos para los mismos (Murphy, 2012, pp. 247-249).

Posted by Eddie Soong on June 29, 2015 at 8:03pm, & Blog, V. (s. f.). *Statistical Modeling steps*. Recuperado 30 de mayo de 2020, de https://www.datasciencecentral.com/profiles/blogs/statistical-modeling-steps

Ruder, S. (2016). An overview of gradient descent optimization algorithms. *CoRR, abs/1609.04747*. http://arxiv.org/abs/1609.04747

Shiflet, A. B., & Shiflet, G. W. (2014). *Introduction to Computational Science: Modeling and Simulation for the Sciences (Second Edition)* (2nd ed.). Princeton University Press.

Tensorflow/probability. (s. f.). GitHub. Recuperado 17 de septiembre de 2019, de https://github.com/tensorflow/probability

8. ESTUDIO DEL SESGO EN EL DISEÑO DE MODELOS DE APRENDIZAJE COMPUTACIONAL: ESTADO DEL ARTE Y TAREAS ESPECÍFICAS

En el capítulo anterior se desarrolló un breve ejemplo usando el modelo de regresión logística bayesiana, recurrente en los siguientes capítulos. Antes de tomar el caso de estudio central para este trabajo, es importante hablar más acerca del estado del arte en cuanto a los temas de sesgo y justicia dentro de modelos de aprendizaje computacional, de una forma más específica que lo desarrollado en los primeros capítulos. También se revisará brevemente la distinción entre las tareas de descubrimiento y prevención, establecida en el capítulo de análisis de la discriminación, en la parte correspondiente a la minería de datos y descubrimiento de conocimiento.

8.1 ESTADO DEL ARTE

Considerando el estado del arte en la materia, el presente trabajo es superficial en cuanto a los métodos desarrollados. Esto no es negativo para el objetivo del trabajo, y en todo caso es conveniente, pues un modelo simple como la regresión logística facilita el análisis de los cambios atribuibles a la metodología bayesiana, como se esbozó en el capítulo anterior.

El estado del arte para la justicia dentro del aprendizaje computacional es complejo y difícil de describir con precisión. Esto se debe a la relativa novedad del tema y su rápido crecimiento en los últimos años; los artículos de investigación en esta área han incrementado de manera exponencial en este periodo (*CS 294: Fairness in Machine Learning*, s/f). Para contextualizar el presente escrito, es mejor referirse al trabajo relacionado que se mencionará al introducir el caso de estudio, pero no debe omitirse el estado actual del campo en general.

Vale la pena mencionar algunas herramientas desarrolladas por compañías importantes en el campo. Google publicó hace menos de un año ML-fairness-gym en GitHub, un proyecto orientado a estudiar el impacto a largo plazo de las medidas tomadas para procurar la justicia en los sistemas de decisiones automatizados, pues en ocasiones dichas medidas pueden tener efectos contraproducentes en términos de justicia en un entorno dinámico a largo plazo; este proyecto ofrece componentes para construir simulaciones que permitan estudiar estas situaciones (*google/ml-fairness-gym*, 2019/2020). Facebook también alega que ha desarrollado una herramienta llamada Fairness Flow para determinar si un algoritmo de aprendizaje computacional tiene prejuicios, pero el código no está publicado como en el caso de ML-fairness-gym (Gershgorn, s/f). Por su parte, IBM publicó en 2018 AI Fairness 360, un kit de herramientas abierto (*open source*) con métricas para revisar la existencia de sesgo indeseable en conjuntos de datos y modelos de aprendizaje computacional y que también provee algoritmos al nivel del estado del arte para mitigar dicho sesgo (*Introducing AI Fairness 360, A Step Towards Trusted AI - IBM Research*, 2018). Finalmente vale la pena mencionar el proyecto UnBias, que comenzó desde 2016, y busca la transparencia en las decisiones automatizadas, no sólo atacando el dilema entre la transparencia en las decisiones y los autores intelectuales de los algoritmos, sino también la problemática de lograr una transparencia comprensible para cualquier persona afectada por el proceso, y no sólo para quienes sepan interpretar el código y los algoritmos empleados (“Our Mission”, 2016).

Más allá de estos proyectos a gran escala, hay una gran cantidad de trabajos que atacan situaciones y casos más específicos, y algunos ya han sido referencias en capítulos pasados. Por mencionar sólo algunos, que fueron consultados durante la elaboración del presente trabajo: en (B Srivastava & F Rossi, s/f) se propone un esquema conceptual para evaluar el sesgo en servicios de inteligencia artificial, independientemente de los desarrolladores de los servicios y de una forma que permite evaluar las implicaciones de usar varios servicios considerando el sesgo de cada uno de ellos, y el sesgo compuesto, de una manera sistematizada. En (Calmon et al., 2017), se desarrolla una formulación probabilística del pre-procesado de datos para reducir el sesgo algorítmico balanceando tres factores: el control de la discriminación, la distorsión de muestras individuales y la preservación de la utilidad. Por su parte, Fish et al. (2016) desarrollan un método para lograr justicia algorítimica modificando la frontera de decisión de los clasificadores, y además proponen una

medida de la justicia, argumentando que permite distinguir entre algoritmos ingenuamente justos y algoritmos con un sentido de justicia más sensato; establecen que un algoritmo puede aparentar ser "justo" con modificaciones ingenuas que no deberían considerarse como soluciones al problema del sesgo.

En cuanto al enfoque en clasificadores bayesianos, también hay varios trabajos que emplean la metodología bayesiana en el contexto de la justicia y el sesgo. Perrone et al. (2020) proponen un método llamado *Fair Bayesian Optimization* (optimización bayesiana justa), el cual es flexible con respecto a los modelos usados, y tiene resultados favorables al optimizar dichos modelos considerando varias nociones de justicia. Dimitrakakis et al. (2017) exploran la toma de decisiones en el contexto del aprendizaje por refuerzo (*reinforcement learning*), proponiendo un método bayesiano para incorporar la incertidumbre en los parámetros de los modelos en las nociones de justicia, obteniendo un rendimiento prometedor conforme se da más énfasis a la justicia de las decisiones como parte de la utilidad de las mismas.

Podría hacerse referencia a otros trabajos en la materia, y se hará cuando se discuta el trabajo relacionado con el caso de estudio central de este escrito, pero lo importante en esta sección es puntualizar las limitaciones de la tesis en el contexto general de la materia y el estado del arte. Primeramente, los modelos presentes en gran parte de los trabajos son más sofisticados que la regresión logística, incluyendo redes neuronales, bosques aleatorios, métodos de ensamble, redes bayesianas, entre otros. En segundo lugar, las nociones de justicia más usadas suelen ir más allá de las medidas de discriminación expuestas por Romei & Ruggieri (2014) que se usan en este trabajo, y en vez de ser cantidades se expresan como condiciones. Algunas de ellas son balance, calibración, paridad estadística, equidad de trato, entre otras (Berk et al., 2017), (Kleinberg et al., 2016). Finalmente, el presente trabajo se distingue de la mayoría en que por lo general el uso de los modelos se enfoca a la tarea de prevención de la discriminación, y en este caso, por la naturaleza del caso de estudio, el enfoque está en la tarea de descubrimiento, estas tareas se plantean en la siguiente sección.

8.2 TAREAS DE PREVENCIÓN Y DESCUBRIMIENTO

Estas tareas fueron descritas en el capítulo de análisis de la discriminación[55]. La prevención consiste en alterar el modelo, ya sea preprocesando datos, modificando el algoritmo de aprendizaje, modificando los parámetros del modelo obtenido, o bien, modificando las predicciones hechas a fin de mantener los resultados en las proporciones deseadas. La gran mayoría de los trabajos mencionados se dedican a esta tarea de prevención, ideando métodos aplicables a los datos o al modelo, para satisfacer diversas condiciones de justicia.

En este trabajo, se esbozó de una manera muy superficial y general el tipo de ventajas que puede ofrecer la metodología bayesiana en la tarea de prevención en el capítulo pasado, donde se probaron modelos de regresión logística sobre datos sintéticos. En ese capítulo, y en el trabajo en general, este es el tipo de ventajas que se exploran, enfocadas en la información disponible y su forma para quien toma las decisiones. Esto contrasta con la mayoría de los trabajos en el tema, donde más bien se comparan rendimientos, considerando definiciones de justicia como condiciones a ser satisfechas o como partes explícitas de las funciones de costo o de utilidad. El camino que se toma en este trabajo se debe a que una parte importante del mismo es la comparación conceptual entre métodos bayesiano y frecuentista, más allá de comparar métricas de rendimiento que dependerían de las nociones de justicia consideradas.

La tarea de descubrimiento será la más importante en este trabajo, pues el caso de estudio se centra en la misma. Las especificidades se expondrán con más detalle en el capítulo correspondiente. Con los múltiples trabajos que se han consultado, parece que el enfoque suele estar en la prevención, pero aún así hay varios trabajos acerca del descubrimiento de la discriminación. Sin embargo, el trabajo dedicado a explorar específicamente el uso de modelos bayesianos en esta

55 Acorde a Romei & Ruggieri (2014).

tarea es relativamente escaso y no suele concentrarse en una comparación con modelos frecuentistas; estos detalles serán revisados al hablar del trabajo relacionado en el planteamiento del caso de estudio.

REFERENCIAS

B Srivastava, & F Rossi. (s/f). *Towards composable bias rating of AI services*.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2017). *Fairness in Criminal Justice Risk Assessments: The State of the Art*.

Calmon, F. P., Wei, D., Ramamurthy, K. N., & Varshney, K. R. (2017). *Optimized Data Pre-Processing for Discrimination Prevention*.

CS 294: Fairness in Machine Learning. (s/f). Recuperado el 11 de junio de 2020, de https://fairmlclass.github.io/1.html#/22

Dimitrakakis, C., Liu, Y., Parkes, D., & Radanovic, G. (2017). *Bayesian fairness*.

Fish, B., Kun, J., & Lelkes, Á. D. (2016). A confidence-based approach for balancing fairness and accuracy. *Proceedings of the 2016 SIAM International Conference on Data Mining*, 144–152.

Gershgorn, D. (s/f). *Facebook says it has a tool to detect bias in its artificial intelligence*. Quartz. Recuperado el 13 de junio de 2020, de https://qz.com/1268520/facebook-says-it-has-a-tool-to-detect-bias-in-its-artificial-intelligence/

Google/ml-fairness-gym. (2020). [Python]. Google. https://github.com/google/ml-fairness-gym (Original work published 2019)

Introducing AI Fairness 360, A Step Towards Trusted AI - IBM Research. (2018, septiembre 19). IBM Research Blog. https://www.ibm.com/blogs/research/2018/09/ai-fairness-360/

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). *Inherent Trade-Offs in the Fair Determination of Risk Scores*.

Our Mission. (2016, julio 8). *UnBias*. https://unbias.wp.horizon.ac.uk/our-mission/

Perrone, V., Donini, M., Kenthapadi, K., & Archambeau, C. (2020). *Fair Bayesian Optimization*.

Romei, A., & Ruggieri, S. (2014). A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review*, *29*(5), 582–638.

9. CASO DE ESTUDIO: PROPUBLICA, COMPAS, Y EL SESGO EN LA PREDICCIÓN DE REINCIDENCIA DELICTIVA

En este capítulo el objetivo es plantear con claridad el caso de estudio sobre el cual se basa este trabajo. Se describirá el problema y se justificará su uso en la realización de esta tesis, para posteriormente hablar del trabajo relacionado que ya se ha hecho. Los resultados y conclusiones de esta tesis se basan en este caso particular, pero la idea es establecer puntos mediante este ejemplo que sean generales para otro tipo de aplicaciones e implementaciones.

9.1 DESCRIPCIÓN DEL PROBLEMA

Una aplicación popular y delicada de los procesos de decisión, automatizados o asistidos por el aprendizaje computacional, es la evaluación de riesgos dentro del sistema penal, particularmente los instrumentos de predicción de reincidencia[56] (RPI por sus siglas en inglés) (Chouldechova, 2017, p. 1). Estos instrumentos son populares en los Estados Unidos debido a los altos índices de encarcelamiento y los problemas secundarios que implican para el país (W. Flores et al., 2016, pp. 2–6).

En los últimos años ha habido debate y polémica alrededor de los RPI, por las alarmantes posibilidades relacionadas con su uso, pues se emplean en decisiones de libertad condicional, en fases previas al juicio, e incluso en las sentencias. Lógicamente, parte de la polémica está basada en la posibilidad de cometer errores de clasificación en este contexto tan sensible, pero la principal preocupación es la injusticia sistemática, es decir, errores de clasificación que afecten de forma consistente a grupos específicos de la población (Chouldechova, 2017, p. 1).

Un ejemplo de esta polémica es un artículo del 2016 de la empresa ProPublica. En él, afirmaron que un RPI llamado COMPAS, desarrollado por la empresa Northpointe, tenía un prejuicio inherente contra la población negra. ProPublica usó los datos de miles de individuos evaluados con COMPAS para desarrollar su análisis, los cuales están disponibles públicamente junto con el código que desarrollaron (escrito en el lenguaje y ambiente estadístico R) (Julia Angwin, 2016), (Larson et al., s/f), (*propublica/compas-analysis*, 2016/2020).

Unos meses después, se publicó otro artículo desacreditando las aseveraciones del trabajo de ProPublica, denunciando errores en su desarrollo, que supuestamente causaron conclusiones erróneas tras malinterpretar los resultados. Dicho escrito presenta sus propios resultados y los autores respaldan la validez de COMPAS[57] (W. Flores et al., 2016).

Ambos estudios presentan sus resultados en términos de la estadística clásica, y ambos usan modelos de regresión logística para estudiar la disparidad racial en este contexto[58]. La experimentación consistirá en desarrollar una regresión logística bayesiana, para estudiar el impacto de la metodología bayesiana en el tipo de conclusiones que pueden obtenerse de los resultados, para ambos estudios. Primero se reproducirán ambos análisis usando Python, que es el lenguaje empleado a lo largo de este trabajo[59].

En este caso, la metodología bayesiana no se aplica a la resolución del problema que el RPI resuelve en sí, es decir, clasificar a los individuos en categorías de riesgo. Más bien se aplica a la

56 En ocasiones también denominados como instrumentos actuariales de evaluación de riesgo (ARAI por sus siglas en inglés) (W. Flores et al., 2016, p. 2).

57 Cabe mencionar que el artículo de ProPublica en sí, responde en varios aspectos a un artículo previo de Northpointe, publicado en 2008 (Brennan et al., 2008). En dicho artículo se evaluó la validez de COMPAS, y se concluyó que la herramienta no tiene prejuicios étnicos (e incluso en esta conclusión ya se respondía a otro estudio del mismo año, que alegaba la existencia de dicho prejuicio en COMPAS antes de que ProPublica hiciera su estudio (Fass et al., 2008)).

58 Ambos estudios contienen una segunda parte dedicada a estudiar la precisión predictiva del modelo, pero no se considerará en este trabajo.

59 ProPublica hizo público su código, que se ejecuta sobre Python, pero en realidad escrito en el lenguaje R, empleando un paquete que permite ejecutar código de R sobre Python. En el segundo caso, los autores no publicaron el código, así que la reproducción es con base en su descripción del procedimiento y resultados.

evaluación de dicho RPI, ya que en este caso, siguiendo el trabajo realizado en ambos estudios, la misma evaluación se basa en un modelo de clasificación.

9.2 JUSTIFICACIÓN

Este caso de estudio es relevante y adecuado para el tema en cuestión por más de una razón. Primero, es un caso donde la justicia algorítmica se traduce de manera concreta en justicia real, pues las injusticias y sesgos en modelos de toma de decisiones para este tipo de aplicación, se convierten inmediatamente en sesgos sistemáticos dentro del sistema legal.

En segundo lugar, el estudio de ProPublica y la respuesta correspondiente se distinguen de la mayoría de los trabajos en el área al usar modelos de clasificación (en este caso de regresión logística) para evaluar el sesgo directamente, interpretando los coeficientes acorde a su significado bajo el modelo de regresión logística (esto es cierto especialmente para el estudio de ProPublica). Esto es conveniente no sólo porque revela con claridad las diferencias entre el método bayesiano y el frecuentista, sino que además permite estudiar la implementación de modelos clasificadores, y a la vez la tarea de evaluación, cuando por lo general los modelos se implementan en la tarea de prevención: es decir, lo usual sería proponer modelos que hagan la misma tarea de clasificación que hace la herramienta COMPAS y den sus propias predicciones en un esquema que busque un balance entre justicia y rendimiento, y no modelos que se ajustan a los datos, incluyendo las predicciones ya hechas por COMPAS, como se da en este caso. Estas cualidades permiten que el presente estudio se enfoque en una perspectiva relativamente inexplorada, que además es idónea para ilustrar las ventajas fundamentales del método bayesiano en términos de la interpretación de los modelos.

Finalmente, como se verá en el trabajo relacionado a continuación, este caso ha sido usado ampliamente en la literatura del tema, pero por lo general los datos se consideran para proponer modelos y analizar su rendimiento en el contexto de la justicia, es decir, las implementaciones buscan llevar a cabo la función de COMPAS mediante diversos conceptos y esquemas. En este caso, el objetivo es estudiar la controversia misma del estudio original y la replica correspondiente, reproduciendo los modelos originales y después implementándolos bajo el esquema bayesiano. Esto permite dar énfasis a las conclusiones que pueden obtenerse de los modelos, en términos del razonamiento humano ante los resultados, desviando este trabajo de la metodología usual que consiste en establecer y comparar métricas. Esto permite estudiar las implicaciones del método bayesiano de manera esencial y general, donde se puede hablar de justicia y rendimiento en términos generales, y no en cuanto a métricas basadas en nociones de justicia bien definidas matemáticamente.

Lo mencionado de ninguna manera significa que este estudio sea superior a la mayoría de los estudios realizados, y de hecho, como se ha sugerido antes, es inferior en muchos aspectos, pues este trabajo es bastante básico en términos de la complejidad de los modelos, y en cuanto a las medidas de discriminación empleadas en comparación a las nociones de justicia que se proponen y desarrollan hoy en día en el área. Lo que sí se busca plantear es que la justificación para las implementaciones desarrolladas, y el caso de estudio mismo, es el enfoque del presente trabajo: estudiar los clasificadores bayesianos en el contexto de la justicia, desde el punto de vista del tipo de respuestas que ofrecen a quien toma las decisiones y saca conclusiones de los datos y resultados obtenidos. Este es un punto que no suele tratarse directamente como se hace en este escrito, y por eso es un tema interesante a pesar de su relativa simpleza.

9.3 TRABAJO RELACIONADO

Como ya se sugirió, los datos del estudio original de ProPublica, son usados ampliamente en trabajos acerca de justicia en el aprendizaje computacional, aunque la mayoría de las veces el estudio original, su análisis y las replicas en su contra, no son tratados. Algunos ejemplos de trabajos que usan dichos datos son: (Calmon et al., 2017), (Dimitrakakis et al., 2017), (Kleinberg et al., 2016), (Chouldechova, 2017), (Perrone et al., 2020).

En cuanto al estudio de la controversia de la herramienta COMPAS entre (Julia Angwin, 2016) y (W. Flores et al., 2016), Kleinberg et al. (2016) estudian y formalizan tres condiciones de justicia presentes en varios debates, para mostrar la imposibilidad de cumplir todas simultáneamente, y este caso de COMPAS es uno de los debates analizados. Por su parte, Chouldechova (2017) analiza los mismos datos de COMPAS para explicar los resultados contradictorios, haciendo la conexión entre la noción psicométrica de la justicia y las tasas de error de clasificación, para después mostrar cómo distintas tasas de error pueden implicar impacto desigual.

En cuanto al efecto de emplear una metodología bayesiana ante una frecuentista en esta tarea, Dimitrakakis et al (2017) y Perrone et al. (2020) proponen métodos bayesianos en el contexto de la justicia, y ambos resuelven la tarea de clasificación de COMPAS. Sin embargo Perrone et al. (2020) simplemente usan los datos del caso de COMPAS para comparar el rendimiento de su método con otros métodos especializados en este tipo de tareas, y no hace claro énfasis en la cualidad bayesiana de su propuesta. Por su parte, Dimitrakakis et al. (2017) sí hacen notar que en su método es importante la manera de incorporar incertidumbre sobre los parámetros, atribuible al método bayesiano, aunque se enfoca al aprendizaje por refuerzo, y no al aprendizaje supervisado como en la mayoría de los trabajos citados y esta tesis. Ninguno de estos dos trabajos estudia la controversia asociada con la implementación original de ProPublica y la replica, y ninguno estudia el descubrimiento y evaluación del sesgo del debate original, como se hace en este trabajo.

Otro tema importante es el uso de métodos y modelos, en particular bayesianos, para detectar injusticias en otros modelos o tareas de clasificación, como es el caso en este trabajo, y en el caso de estudio a tratar, tanto para el estudio original como para la réplica. McNair (2018) presenta un trabajo interesante, donde explora métodos frecuentistas y bayesianos para evaluar la justicia en modelos de toma de decisión de aprendizaje computacional en un contexto médico. En dicho trabajo no se plantea explícitamente la comparación entre métodos, simplemente se aplican dos métodos frecuentistas (prueba de Cochran-Mantel-Haenszel y regresión beta) y uno bayesiano (promedio de modelos bayesianos o *Bayesian model averaging*) para la evaluación de injusticias. Las respuestas obtenidas por los tres modelos son compatibles, sin embargo, el autor concluye que el método bayesiano empleado ofrece varias ventajas sobre sus pares frecuentistas, entre ellas la interpretación comprensible de manera inmediata, y que evita las suposiciones al calcular la significación estadística en la diferencia de tratamientos, ventajas que ya se esbozaron en el capítulo que discute la metodología bayesiana.

Finalmente, trabajo relacionado tanto al uso de modelos bayesianos en la evaluación de la injusticia como al debate original alrededor de COMPAS, se expone en (*Causal Bayesian Networks*, s/f). Este trabajo se compone de dos artículos: en el primero, se emplean redes causales bayesianas como una herramienta visual, para estudiar la injusticia inherente a los datos usados por un modelo y las relaciones potencialmente injustas entre los atributos (Chiappa & Isaac, 2019). En ese mismo artículo se considera el debate de la herramienta COMPAS, concluyendo que el debate en general no da suficiente consideración a los patrones de injusticia subyacentes en los datos. En un segundo artículo, el mismo método se emplea en el contexto de técnicas de inferencia contrafactuales, como una herramienta para cuantificar y mitigar injusticias en los datos (Chiappa & Gillam, 2018). Aunque en estos trabajos se emplean métodos bayesianos para estudiar el caso de estudio, no se emplean modelos de clasificación directamente como en el estudio original, la réplica, y la presente tesis. Al no usar dichos modelos, tampoco se da el enfoque a los cambios particulares de cambiar la implementación original por su equivalente bayesiana, como se desarrollará en los próximos capítulos. A pesar de ello es un antecedente relevante, y de los trabajos más similares en esencia a esta tesis, y por ende debe mencionarse.

REFERENCIAS

Brennan, T., Dieterich, W., & Ehret, B. (2008). Evaluating the Predictive Validity of the Compas Risk and Needs Assessment System. *Criminal Justice and Behavior*, *36*(1), 21–40. https://doi.org/10.1177/0093854808326545

Calmon, F. P., Wei, D., Ramamurthy, K. N., & Varshney, K. R. (2017). *Optimized Data Pre-Processing for Discrimination Prevention*.

Causal Bayesian Networks: A flexible tool to enable fairer machine learning. (s/f). Deepmind. Recuperado el 15 de junio de 2020, de /blog/article/Causal_Bayesian_Networks

Chiappa, S., & Gillam, T. P. S. (2018). *Path-Specific Counterfactual Fairness*.

Chiappa, S., & Isaac, W. S. (2019). A Causal Bayesian Networks Viewpoint on Fairness. *IFIP Advances in Information and Communication Technology*, 3–20. https://doi.org/10.1007/978-3-030-16744-8_1

Chouldechova, A. (2017). *Fair prediction with disparate impact: A study of bias in recidivism prediction instruments*.

Dimitrakakis, C., Liu, Y., Parkes, D., & Radanovic, G. (2017). *Bayesian fairness*.

Fass, T., Heilbrun, K., DeMatteo, D., & Fretz, R. (2008). The LSI-R and the CompasValidation Data on Two Risk-Needs Tools. *Criminal Justice and Behavior - CRIM JUSTICE BEHAV*, *35*, 1095–1108. https://doi.org/10.1177/0093854808320497

Julia Angwin, J. L. (2016, mayo 23). *Machine Bias* [Text/html]. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). *Inherent Trade-Offs in the Fair Determination of Risk Scores*.

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (s/f). *How We Analyzed the COMPAS Recidivism Algorithm*. ProPublica. Recuperado el 2 de mayo de 2020, de https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

McNair, D. S. (2018). Preventing Disparities: Bayesian and Frequentist Methods for Assessing Fairness in Machine-Learning Decision-Support Models. En M. S. F. Nezhad (Ed.), *New Insights into Bayesian Inference*. IntechOpen. https://doi.org/10.5772/intechopen.73176

Perrone, V., Donini, M., Kenthapadi, K., & Archambeau, C. (2020). *Fair Bayesian Optimization*.

Propublica/compas-analysis. (2020). [Jupyter Notebook]. ProPublica. https://github.com/propublica/compas-analysis (Original work published 2016)

W. Flores, A., Bechtel, K., & Lowenkamp, C. (2016). False Positives, False Negatives, and False Analyses: A Rejoinder to “Machine Bias: There’s Software Used Across the Country to Predict Future Criminals. And it’s Biased Against Blacks.” *Federal probation*, *80*.

10. ANÁLISIS DE PROPUBLICA

En este capítulo se estudia el análisis original de ProPublica (Larson et al., s. f.). Primero se implementa una reproducción del trabajo original siguiendo su análisis y código, usando los mismos datos, los cuales también publicaron (*propublica/compas-analysis*, 2016/2020); se presentan los resultados obtenidos de la reproducción, comparando con los originales, y también se discute el análisis de ProPublica para dichos resultados. Posteriormente se plantea el diseño de los modelos bayesianos a implementar, de forma conceptual, representando también los modelos originales. Finalmente se presentan y se discuten los resultados para la implementación bayesiana, comparando con los resultados y conclusiones originales.

10.1 REPRODUCCIÓN

El código de ProPublica en realidad está escrito en Python, pero empleando un módulo llamado rpy2 que permite programar con el lenguaje R sobre el entorno de Python (*Documentation for rpy2 — rpy2 3.3.1 documentation*, s. f.). La reproducción se hizo directamente en Python, empleando varios módulos, donde sobresalen Pandas y TensorFlow, ya mencionados anteriormente. Los resultados obtenidos concuerdan con los publicados en el artículo original.

Preprocesamiento de Datos

Los datos adquiridos por ProPublica vienen originalmente de parte de las autoridades del condado de Broward en Florida, y corresponden a todos los detenidos evaluados con COMPAS entre 2013 y 2014. El análisis inicia con un preprocesamiento de los datos, con el objetivo de emplear sólo los datos de personas evaluadas antes de ser enjuiciadas (la herramienta también se usa en otras etapas del sistema penal).

El preprocesamiento consiste en remover registros con alguna de estas cuatro características: tener más de 30 días de diferencia entre fecha de arresto y fecha de cargo; no corresponder a un caso con evaluación de COMPAS; corresponder a ofensas que no resultaron en encarcelamiento; ser casos donde la reincidencia tardó más de dos años en ocurrir, o donde tras la liberación no pasaron por lo menos dos años con el individuo fuera de cualquier institución correccional.

Una vez que se han filtrado los datos bajo los criterios mencionados, queda un total de 6172 elementos (antes de filtrar el total era 7214). Estos datos son explorados para observar varios aspectos de los mismos:

Correlación entre puntaje COMPAS y tiempo de encarcelamiento: tras definir un atributo temporal para el tiempo de encarcelamiento, se usa una función de R para determinar su correlación con el puntaje COMPAS y se halla una ligera correlación de 0.2073297. Al reproducir este proceso, usando una función equivalente de la librería Pandas, se halla un valor cercano de 0.2074120.

Frecuencias en la población: posteriormente se agrupan los elementos por raza, género, puntaje COMPAS y reincidencia, o por combinaciones de estos atributos. Los resultados son consistentes con los expuestos en (*propublica/compas-analysis*, 2016/2020).

Edad	*menos de 25*	*25 - 45*	*más de 45*
Elementos en la población	1347	3532	1293

Tabla 10.1.1. Elementos de la población agrupados por edad.

Raza	*Afroamericana*	*Asiática*	*Caucásica*	*Hispana*	*Nativo Americana*	*Otra*
Elementos en la población	3175	31	2103	509	11	343
Porcentaje de la población total	51.44%	0.50%	34.07%	8.25%	0.18%	5.56%

Tabla 10.1.2. Elementos de la población agrupados por raza.

Puntaje COMPAS	*Bajo*	*Medio*	*Alto*
Elementos en la población	3421	1607	1144

Tabla 10.1.3. Elementos de la población agrupados por puntaje COMPAS.

	Raza					
Género	*Afroamericana*	*Asiática*	*Caucásica*	*Hispana*	*Nativo americana*	*Otra*
Femenino	549	2	482	82	2	58
Masculino	2626	29	1621	427	9	285

Tabla 10.1.4. Elementos de la población agrupados por raza y género.

Género	*Femenino*	*Masculino*
Elementos en la población	1175	4997
Porcentaje de la población total	19.04%	80.96%

Tabla 10.1.5. Elementos de la población agrupados por género.

Elementos que reincidieron en un periodo de dos años en la población	2809
Porcentaje de la población total	45.51%

Tabla 10.1.6. Elementos de la población agrupados por reincidencia.

El puntaje de COMPAS está en una escala del 1 al 10, y un mayor puntaje significa un mayor riesgo de reincidencia; los puntajes pueden representar riesgo bajo (1 a 4), medio (5 a 7) o alto (8 a 10). Tras el preprocesamiento, otro análisis inicial de los datos muestra la distribución de puntajes de las poblaciones para razas blanca (caucásica) y negra (afroamericana). Ya se ha visto que hay otras razas, pero el enfoque principal de ProPublica siempre fue comparar estos dos grupos (que además son los dos más numerosos):

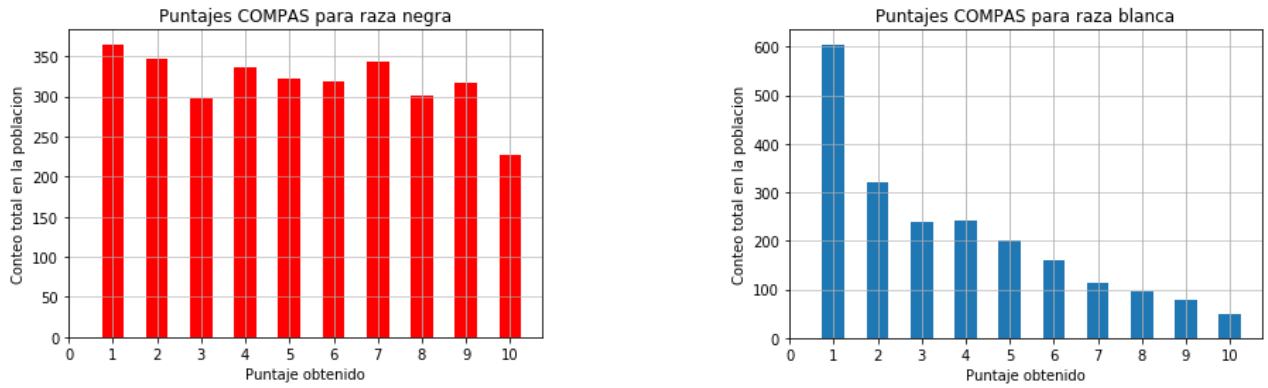


Figura. 10. 1. Miembros en las poblaciones de raza negra y blanca para cada puntaje COMPAS.

Como se puede observar en la figura 10.1, en la población de raza negra hay una distribución mucho más balanceada, y para la población blanca la gran mayoría se clasifica en la categoría de bajo riesgo, y además hay una evidente tendencia decreciente de población con el aumento del puntaje COMPAS. Otro punto que debe tomarse en cuenta es las escasez de elementos en las razas asiática y nativo americana, que será relevante en análisis posteriores de los resultados.

En el estudio original, ProPublica también incluye una tabla donde muestra los conteos de cada puntaje para cada raza, en este caso sólo se presentan los gráficos que contienen dicha información para las razas caucásica y afroamericana, pues el análisis se centra en ellas (como se verá más adelante, la replica de W. Flores et al. (2016) ni siquiera considera otras razas en su estudio). De cualquier forma, esta tabla se presenta en el anexo para resultados adicionales.

El Modelo Clasificador

La segunda fase del proceso estudia el prejuicio racial de COMPAS mediante un modelo de regresión logística. Las categorías en los datos originales son más de cincuenta, pero en el estudio de ProPublica se usan algunos atributos y factores definidos a partir de éstos, que se emplean como las únicas covariables para predecir el puntaje obtenido. Los factores son atributos categóricos y discretos, definidos a partir de varios atributos, y posteriormente codificados (más detalles a continuación). Los atributos originales empleados (conteo de antecedentes y reincidencia en dos años) se mantienen como están en los datos originales. Estos factores y atributos son:

crime_factor (factor de crimen): clasifica a los individuos acorde al tipo de agresión cometida. Puede ser *misdemeanor* (delito menor) o *felony* (delito grave).

age_factor (factor de edad): clasifica a los individuos en tres rangos de edad. Menores de 25, de 25 a 40 años, y mayores de 40.

race_factor (factor de raza): clasifica acorde a la raza de los individuos, y los valores posibles concuerdan con los de los datos originales (caucásico, afroamericano, asiático, hispano, nativo americano, otro).

gender_factor (factor de género): clasifica a los individuos en géneros masculino y femenino.

priors_count (conteo de antecedentes): es un número entero que representa el número de antecedentes delictivos de los individuos.

two_year_recid (reincidencia en dos años): es una variable binaria que indica si el individuo reincidió en los dos años posteriores a la asignación de su puntaje COMPAS.

score_factor (factor de puntaje): clasifica a los individuos acorde al rango de puntaje COMPAS. Como ya se mencionó, numéricamente el puntaje toma valores de 1 a 10, pero a su vez se divide en categorías de riesgo baja (1 a 3), media (4 a 7), y alta (8 a 10). Este factor divide sólo en dos categorías de puntaje: alto y bajo; para conseguirlo, se clasifican todos los puntajes medios como altos.[60]

60 Esta decisión en el análisis de ProPublica fue muy criticada por la respuesta de W. Flores et al. (2016), donde se sugiere que tendría que analizarse el efecto de unir puntajes medios y bajos en vez de medios y altos (esta posibilidad se explora en el anexo de resultados adicionales). Sin embargo la decisión no fue del todo injustificada, ya que se basa en la guía emitida por Northpointe para usuarios de COMPAS, donde se menciona que los puntajes medios y altos son de mayor interés para las agencias de supervisión (equivant, 2019, p. 24).

Una vez establecidos estos factores, se procede a establecer el modelo de regresión logística. La clasificación se hace con el factor de puntaje como salida, pues el objetivo es observar cómo influyen las variables explicativas al determinar si un individuo es clasificado en riesgo alto o bajo.

Haciendo un repaso de capítulos anteriores, en un modelo de regresión logística se modela la probabilidad de éxito de una distribución de Bernoulli. Esto se hace a través de una función sigmoidea sobre los datos de entrada o variables explicativas (**x**). La respuesta (*y*) es entonces la probabilidad de éxito. El objetivo es determinar los pesos (vector **w**) y la constante de sesgo (escalar *b*) que mejor expresen la relación no lineal entre entrada y salida (y que mejor expliquen los datos observados). La función del modelo (*f*) para aproximar la salida es entonces:

$$y \simeq f(\mathbf{w}, \mathbf{x}) = \text{sigm}(\mathbf{wx} + b) = \frac{1}{1+e^{-(\mathbf{wx}+b)}}$$

En este caso, como ya se mencionó, la variable de salida sería el factor de puntaje, que es el puntaje COMPAS ajustado para ser binario (alto o bajo). La probabilidad de éxito será tomada como la de ser clasificado en alto riesgo. Es decir, numéricamente, el alto riesgo representa un 1 y el bajo riesgo un 0.

Las variables de entrada que componen al vector **x** son las dos variables explicativas presentes en los datos originales, *priors_count* (conteo de antecedentes) y *two_year_recid* (reincidencia en dos años), y todos los factores ya mencionados, a excepción del factor de puntaje (es decir, los factores de crimen, edad, raza, y género).

Las variables de antecedentes y reincidencia se usan sin modificarlas, mientras que los llamados factores emplean la codificación denominada "one-hot encoding", que se usa para dar un significado numérico adecuado a variables categóricas. Esta codificación consiste en tomar alguna variable categórica con *n* categorías y convertirla en un vector de *n* o *n -1* valores binarios (*MLPR w3a - Machine Learning and Pattern Recognition*, s. f.). Esto quiere decir que por ejemplo, el factor de raza, en vez de ser representado como una variable que puede tomar uno de los 6 valores discretos posibles para la raza de cada individuo, es representado por 5 variables binarias, una para cada raza, menos la caucásica. De esta forma, un individuo de cualquier otra raza tendrá un valor de 1 en la variable binaria de su raza, y 0 en todas las demás; del mismo modo, si tiene 0 en todas las variables binarias significa que es de raza caucásica. En todo caso podría haber una sexta variable binaria para la raza caucásica, pero el estudio de ProPublica lo estructuró así para tener una serie de características de referencia, que serán relevantes al interpretar el modelo entrenado para las conclusiones (esto se aclara más adelante).

Para obtener los valores de los pesos (**w**) y la constante de sesgo (*b*), en el estudio de ProPublica se empleó la función "glm.fit" de R (*glm.fit function | R Documentation*, s. f.), (*R: Fitting Generalized Linear Models*, s. f.). Para la reproducción en Python, se empleó la función "tfp.glm.fit" del módulo TensorFlow Probability (*Tfp.Glm.Fit | TensorFlow Probability*, s. f.), (*Tensorflow/Probability*, s. f.). Ambas funciones usan un método basado en maximizar la verosimilitud (*maximum likelihood*) llamado *Fisher scoring*[61]. Sin entrar en detalles, los resultados obtenidos en la reproducción fueron prácticamente idénticos a los originales de ProPublica (usando la misma precisión numérica de su análisis). Los resultados son los siguientes:

61 Este método se toma en la documentación como sinónimo del método de mínimos cuadrados iterativos ponderados (IWLS por sus siglas en inglés), pues éste último es una formulación del primero. Al maximizar la verosimilitud logarítmica (*log*-likelihood), el método involucra la práctica habitual de minimizar la verosimilitud logarítmica negativa (*negative log-likelihood* o NLL). Más información acerca de estos detalles, y el método en general, se puede revisar en (Dutang, 2017).

Coeficientes de Regresión Logística (componentes del vector **w** y valor de *b*)		
Atributo correspondiente en el vector **x**	*Resultados de ProPublica*	*Resultados de la reproducción*
Ninguno (el coeficiente es la constante *b*)	-1.52554	-1.52554
Factor de género: femenino (*gender_factor*)	0.22127	0.22127
Factor de edad: mayor de 45 (*age_factor*)	-1.35563	-1.35563
Factor de edad: menor de 25 (*age_factor*)	1.30839	1.30839
Factor de raza: afroamericano (*race_factor*)	0.47721	0.47721
Factor de raza: asiático (*race_factor*)	-0.25441	-0.25441
Factor de raza: hispano (*race_factor*)	-0.42839	-0.42839
Factor de raza: nativo americano (*race_factor*)	1.39421	1.39421
Factor de raza: otro (*race_factor*)	-0.82635	-0.82635
Conteo de antecedentes (*priors_count*)	0.26895	0.26895
Factor de crimen: delito menor (*crime_factor*)	-0.31124	-0.31124
Reincidencia en dos años (*two_year_recid*)	0.68586	0.68586

Tabla 10.2. Coeficientes obtenidos en la reproducción y en el estudio original.

En la tabla se puede observar el efecto de la codificación "one-hot", pues para algunos factores de entrada, hay más de una variable representándolos[62]: se tiene una variable para cada valor posible del factor, menos uno. El valor que se omite se puede deducir de los valores que sí se emplearon en la regresión como variables: en el factor de género se omite el género masculino, en el factor de raza, la raza caucásica, en el factor de edad, el rango entre 25 y 45 años, y en el factor de crimen, los delitos mayores. El significado de esto es que la constante *b* representa a los individuos que están en ceros con respecto a todas las categorías de entrada usadas, es decir, representa a individuos de género masculino, raza caucásica, de entre 25 y 45 años, que cometieran un delito mayor, no tuvieran ningún antecedente y no reincidieran en dos años.

El valor numérico de *b* mezcla varias características y no tiene un significado claro por sí solo, pero representa el efecto que todas estas características en conjunto tienen sobre la probabilidad de ser clasificado en alto riesgo por COMPAS. Lo interesante son los demás coeficientes, que representan el efecto de características específicas (género femenino, raza afroamericana, antecedentes, etc.), y se puede observar el efecto aislado con el cual éstas contribuyen al resultado (clasificación de COMPAS) bajo este modelo de regresión logística.

Las aseveraciones del trabajo de ProPublica en relación con el modelo de regresión logística, están basadas en la medida de riesgo relativo (RR), mencionada en el capítulo referente al análisis de la discriminación. En dicho capítulo las diferentes medidas se describieron en función de

62 Notablemente, para el factor de raza, hay cinco variables binarias, cada una representando a una raza específica. El factor de edad es el único, además del factor de raza, representado con más de una variable.

probabilidades basadas en frecuencias relativas, describiendo cuántos elementos de cada grupo recibían un beneficio y a cuántos se les negaba. En este caso las probabilidades están basadas directamente en la distribución de Bernoulli descrita por el modelo de regresión logística (como se mencionó en el capítulo de clasificadores y en esta sección). Una diferencia en cuanto a la interpretación del problema, es que en este caso no se usa la probabilidad de no recibir un beneficio, sino directamente la probabilidad de recibir un resultado indeseable (la clasificación de alto riesgo en COMPAS).

Nuevamente se obtuvieron resultados equivalentes a los del trabajo original, pero además del RR usado por ProPublica, se obtuvieron otras medidas relevantes[63]:

	Reproducción				**ProPublica**
	RR	*RD*	*RC*	*OR*	*RR*
Raza negra	1.452838	0.080898	0.901506	1.611567	1.452841
Género Femenino	1.194795	0.0348	0.957631	1.247656	1.194798
Edad menor a 25 años	2.49612	0.267278	0.674588	3.700214	2.49612

Tabla. 10.3. Medidas de discriminación considerando distintos grupos protegidos.

Con base en la medida de riesgo relativo (RR), el estudio de ProPublica propuso lo siguiente: los elementos de raza negra tienen una probabilidad 45% mayor de ser clasificados con un puntaje más alto, los elementos de género femenino tienen una probabilidad 19.4% mayor, y los elementos con menos de 25 años de edad tienen una probabilidad 2.5 veces mayor. Específicamente, cuando se dice que los elementos con alguna característica tienen cierto mayor porcentaje de probabilidad de ser clasificados en alto riesgo, en realidad se refiere a lo siguiente: manteniendo todas las demás características fijas dentro del grupo de referencia no protegido[64], el modificar la característica en cuestión resulta en una probabilidad *p*% mayor de tener una clasificación de alto riesgo.

Reincidencia Violenta

De la misma forma, se obtienen los resultados para el caso de reincidencia violenta, que considera casos donde el puntaje COMPAS predice únicamente reincidencia donde los delitos son violentos[65]. El código empleado es el mismo, simplemente se usan datos diferentes (ProPublica da acceso a ambos conjuntos de datos). Del mismo modo que en la reincidencia general, se filtran los datos, resultando en un total de 4020 registros de 4743 disponibles. Siguiendo el mismo procedimiento, se exploran los datos y se presentan los coeficientes de regresión logística y las medidas de la discriminación, tanto en la reproducción como en el estudio original. Ya que los detalles del procedimiento fueron cubiertos en la sección anterior, aquí se presentan directamente los resultados.

correlación entre puntaje COMPAS y tiempo de encarcelamiento: para el puntaje COMPAS para reincidencia violenta, aún hay correlación positiva pero tiene un valor de 0.1645568, menor al valor obtenido para el caso de reincidencia general.

63 Como ya se mencionó en el capítulo de análisis de la discriminación, el RR es la medida usada por la Unión Europea al comparar grupos. Como referencia también se calcularon la diferencia de riesgo (RD), la tasa de selección (RC), y el radio de posibilidades (OR), medidas empleadas por las autoridades del Reino Unido, las autoridades de los Estados Unidos, e investigadores legales, respectivamente (Romei & Ruggieri, 2014, p. 9).

64 Es decir, las características implícitas de un elemento que tiene ceros en todas las variables binarias en cuestión, ya mencionadas antes. En este caso eso significa un elemento de raza blanca, género masculino, de entre 25 y 45 años de edad, que cometió un delito mayor, no tiene antecedentes, y no reincidió.

65 Violentos acorde a la definición del FBI (*Violent Crime*, s. f.).

frecuencias en la población: nuevamente se obtienen las frecuencias en la población, aunque cabe mencionar que en el análisis de ProPublica se omiten algunos de los puntos presentados para el caso de reincidencia general, incluyendo la correlación recién obtenida[66]. En este trabajo se presentan para la reincidencia violenta todos los puntos explorados para reincidencia general.

Edad	*menos de 25*	*25 - 45*	*más de 45*
Elementos en la población	766	2300	954

Tabla 10.4.1. Elementos de la población agrupados por edad (reincidencia violenta).

Raza	*Afroamericana*	*Asiática*	*Caucásica*	*Hispana*	*Nativo Americana*	*Otra*
Elementos en la población	1918	26	1459	355	7	255
Porcentaje de la población total	47.71%	0.65%	36.29%	8.83%	0.17%	6.34%

Tabla 10.4.2. Elementos de la población agrupados por raza (reincidencia violenta).

Puntaje COMPAS	*Bajo*	*Medio*	*Alto*
Elementos en la población	2913	828	279

Tabla 10.4.3. Elementos de la población agrupados por puntaje COMPAS (reincidencia violenta).

	Raza					
Género	*Afroamericana*	*Asiática*	*Caucásica*	*Hispana*	*Nativo americana*	*Otra*
Femenino	393	1	336	61	0	50
Masculino	1525	25	1123	294	7	205

Tabla 10.4.4. Elementos de la población agrupados por raza y género (reincidencia violenta).

Género	*Femenino*	*Masculino*
Elementos en la población	841	3179
Porcentaje de la población total	20.92%	79.08%

Tabla 10.4.5. Elementos de la población agrupados por género (reincidencia violenta).

Elementos que reincidieron en un periodo de dos años en la población	652
Porcentaje de la población total	16.22%

Tabla 10.4.6. Elementos de la población agrupados por reincidencia (usando los datos de reincidencia violenta).

66 Como se mencionó en la sección anterior, se omitió la tabla mostrando el número de elementos de cada raza para cada puntaje COMPAS, es decir, la información representada en los gráficos de la figura 10.1, pero para todas las razas (no sólo afroamericana y caucásica). Esta tabla también se omite para el caso de reincidencia violenta en este trabajo. Estas tablas se muestran en el anexo de este escrito para resultados adicionales, y sus datos son congruentes con los presentados por ProPublica (*propublica/compas-analysis*, 2016/2020).

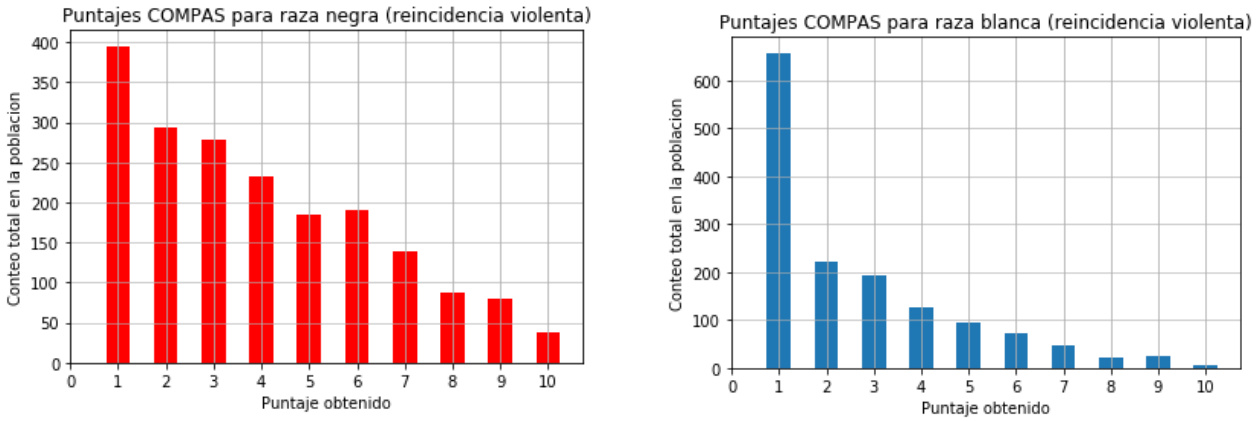


Figura. 10.2. Miembros en las poblaciones de raza negra y blanca para cada puntaje COMPAS considerando sólo crímenes violentos.

Coeficientes de Regresión Logística (componentes del vector **w** y valor de *b*)		
Atributo correspondiente en el vector **x**	*Resultados de ProPublica*	*Resultados de la reproducción*
Ninguno (el coeficiente es la constante *b*)	-2.24274	-2.24273
Factor de género: femenino (*gender_factor*)	-0.7289	-0.7289
Factor de edad: mayor de 45 (*age_factor*)	-1.74208	-1.74208
Factor de edad: menor de 25 (*age_factor*)	3.14591	3.14591
Factor de raza: afroamericano (*race_factor*)	0.65893	0.65893
Factor de raza: asiático (*race_factor*)	-0.98521	-0.98521
Factor de raza: hispano (*race_factor*)	-0.06416	-0.06416
Factor de raza: nativo americano (*race_factor*)	0.44793	0.44793
Factor de raza: otro (*race_factor*)	-0.20543	-0.20543
Conteo de antecedentes (*priors_count*)	0.13764	0.13764
Factor de crimen: delito menor (*crime_factor*)	-0.16367	-0.16367
Reincidencia en dos años (*two_year_recid*)	0.93448	0.93448

Tabla 10.5. Coeficientes obtenidos en la reproducción y en el estudio original (reincidencia violenta).

	Reproducción				**ProPublica**
	RR	*RD*	*RC*	*OR*	*RR*
Raza negra	1.773923	0.074280	0.917834	1.932727	1.773921
Género Femenino	0.507655	1.052271	-0.047254	0.482437	N/A
Edad menor a 25 años	7.414208	0.615623	0.319018	23.240739	7.41424

Tabla 10.6. Medidas de discriminación considerando distintos grupos protegidos (reincidencia violenta).

Nuevamente los resultados de la implementación concuerdan con los resultados de ProPublica, con pequeñas diferencias en los últimos decimales en algunos datos. Para la distribución de puntajes se puede notar que en ambos casos hay una tendencia similar, pero en general los individuos de raza blanca reciben menores puntajes nuevamente, y entre ellos los puntajes altos son más raros.

En cuanto a los coeficientes de regresión logística y las medidas de discriminación obtenidas a partir de éstos, nuevamente los autores de ProPublica emplearon el riesgo relativo (RR) para inferir que los elementos de raza negra tienen una probabilidad 77.3%[67] mayor de recibir un puntaje alto, mientras que elementos de la población joven (menor a 25 años) tienen una probabilidad 7.4 veces mayor de obtener un puntaje alto[68]. Para la reincidencia violenta, ProPublica no obtuvo el riesgo relativo en el caso de elementos de género femenino, pero con base en los resultados de la implementación, la inferencia correspondiente sería que dichos elementos tienen sólo poco más de la mitad de la probabilidad de recibir un puntaje alto comparados a los de género masculino (una probabilidad 49.2% menor).

10.2 DISEÑO DE MODELOS

Ya que se reprodujo el análisis original, el siguiente punto es establecer los modelos que se implementarán para el caso bayesiano. Normalmente este proceso involucra considerar varias posibilidades acorde a los datos y a la recolección de los mismos, pero dado el enfoque de este trabajo, el proceso se simplifica en gran manera, pues se usarán modelos de regresión logística con las mismas variables que en el estudio original. En este caso son dos modelos: el modelo para reincidencia general y el modelo para casos de reincidencia violenta. El proceso también se agiliza dado que ya se expusieron diversas decisiones adicionales para los modelos en el capítulo de clasificadores (capítulo 5) y en el ejemplo con datos sintéticos (capítulo 7).

El objetivo de esta sección es ilustrar con mayor claridad el desarrollo implementado y los modelos usados. Aunque la mayoría de los detalles del proceso ya fueron discutidos y justificados en capítulos anteriores, y en la sección pasada referente a la reproducción del trabajo original es bueno tener un diagrama conceptual que resuma los diversos conceptos y características.

Preprocesamiento

El preprocesamiento en este caso no depende de usar o no el método bayesiano, por lo que la etapa no se repite al implementarlo. A continuación se presenta el esquema para esta etapa, que es igual en ambos métodos.

67 Esta cifra debería redondearse a 77.4%, pero en el estudio original se presenta como 77.3% (*propublica/compas-analysis*, 2016/2020).

68 Nótese que en la presentación del análisis de ProPublica, esta cifra se presenta erróneamente como "6.4 veces mayor" (Larson et al., s. f.). Error que no está presente en la publicación completa del código y el análisis en GitHub (*propublica/compas-analysis*, 2016/2020).

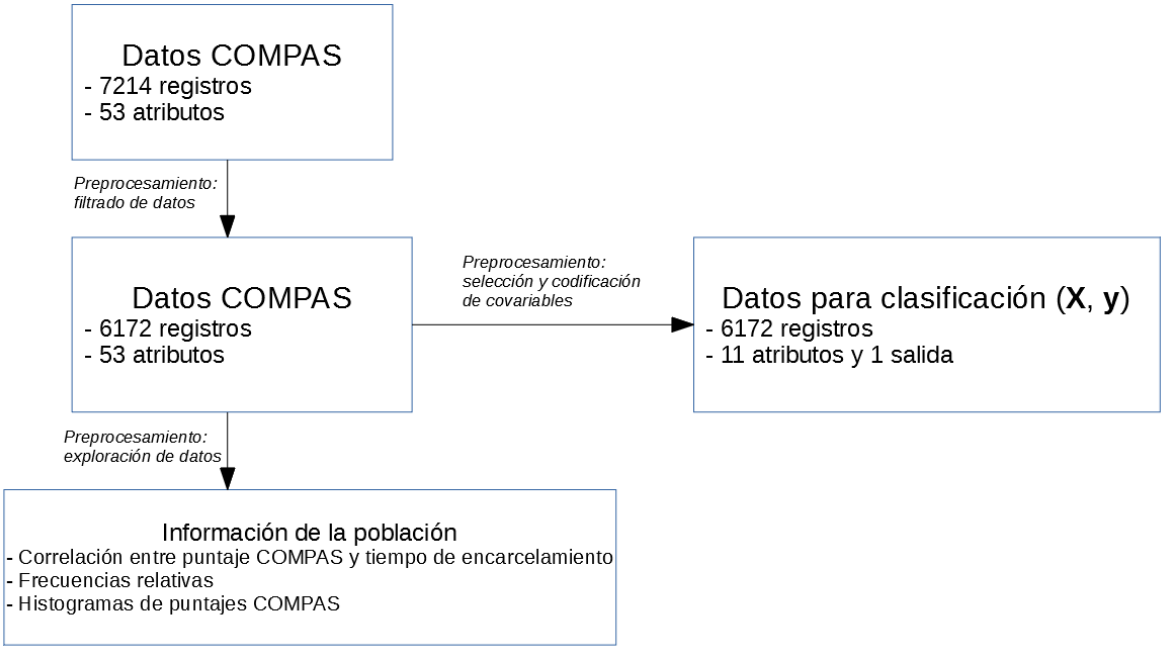


Figura. 10.3.1. Esquema para el preprocesamiento de datos.

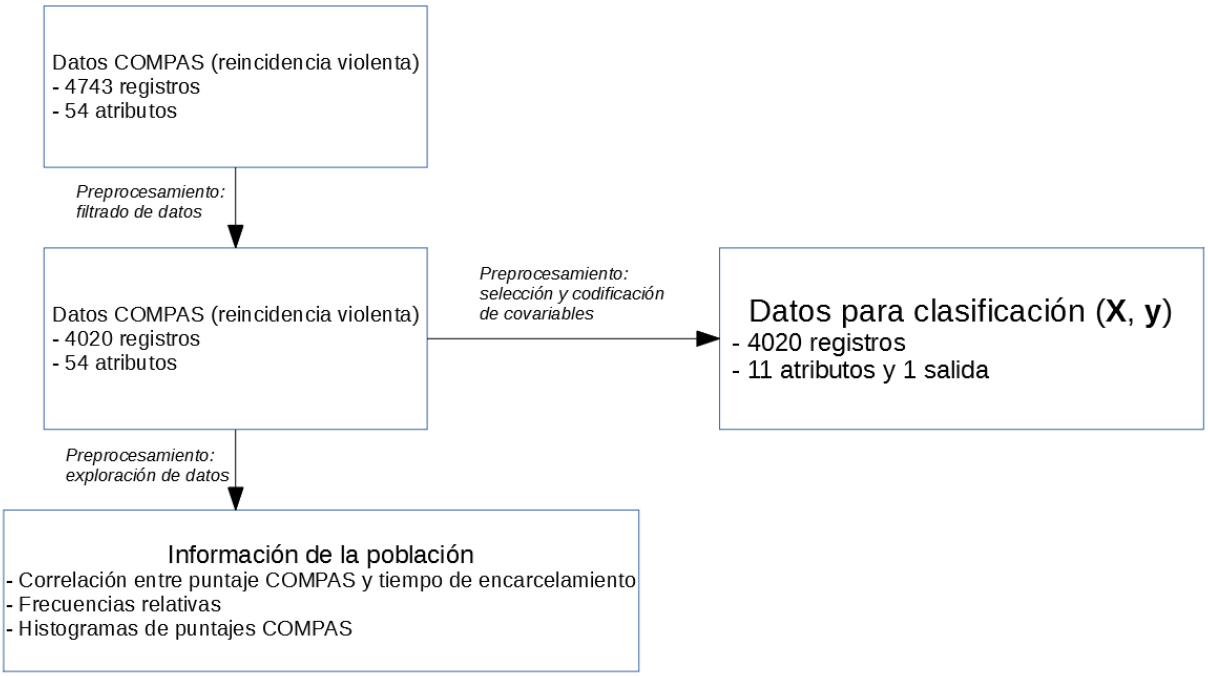


Figura. 10.3.2. Esquema para el preprocesamiento de datos (reincidencia violenta).

En ambos casos, el filtrado inicial descarta registros con inconsistencias, y una vez que se filtran los datos, se puede proceder, por un lado con la exploración de los mismos, y por el otro con la selección de variables y codificación de factores que dan como resultado los datos de la tarea de clasificación, definidos en términos de entradas **X** y salidas **y**. Algo que no se mencionó antes, porque no tiene relevancia para los modelos o el análisis general, es que los datos de reincidencia violenta tienen un atributo más que los de reincidencia general, pero es indistinto porque las 11 variables relevantes en el estudio están presentes en ambos casos.

Modelado

En este caso, la decisión de modelo se reduce a elegir entre el método bayesiano y el frecuentista, pues ya se estableció que el clasificador será la regresión logística, y del mismo modo se establecieron los detalles de su implementación. A continuación se representa esta etapa gráficamente.

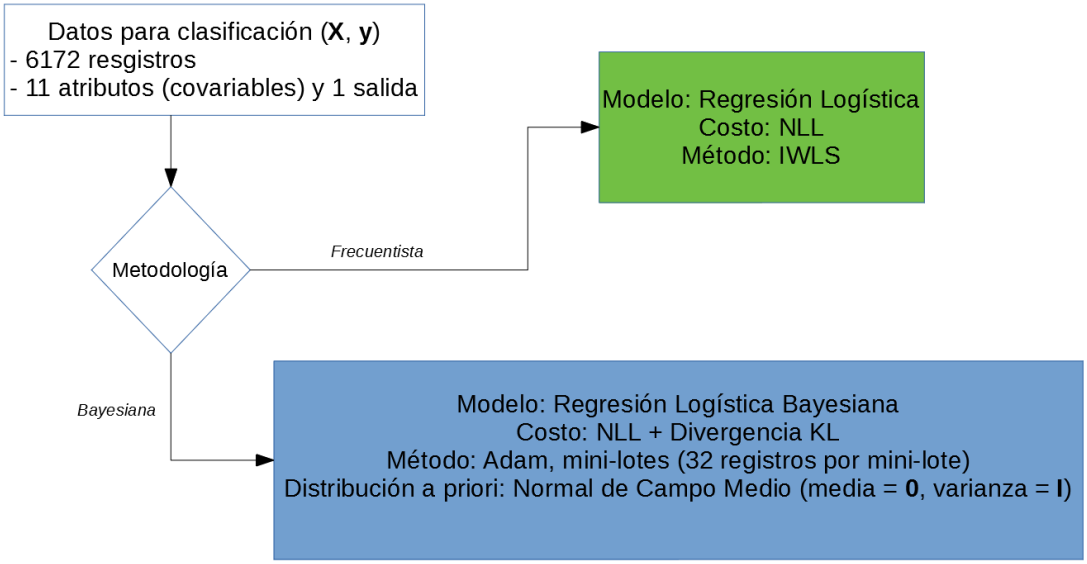


Figura. 10.4.1. Esquema para la selección e implementación de metodologías y modelos.

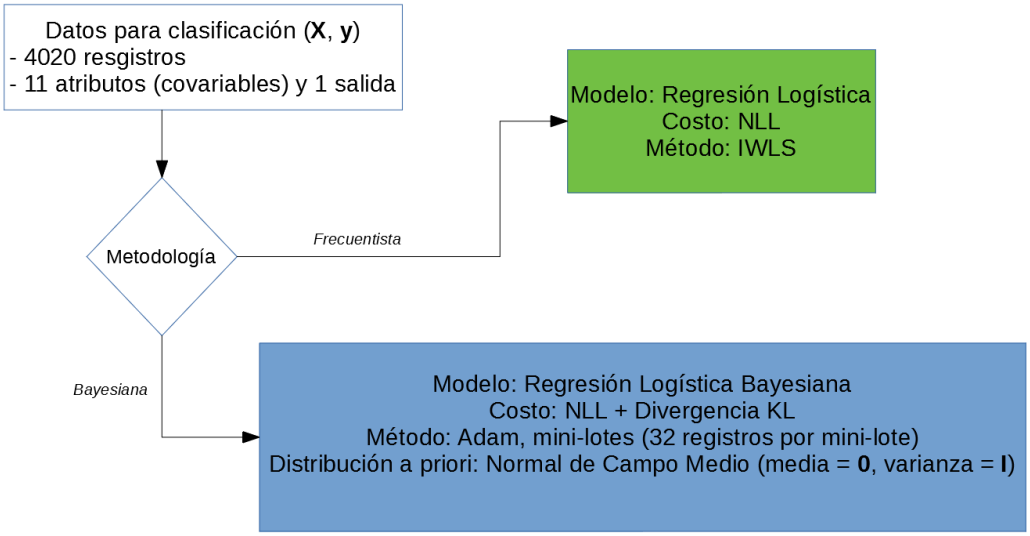


Figura. 10.4.2. Esquema para la selección e implementación de metodologías y modelos (reincidencia violenta).

La única diferencia entre los casos para reincidencia general y reincidencia violenta, son los datos de entrada. Como ya se mencionó, normalmente elegir el modelo clasificador es parte del proceso. En este caso hasta cierto punto la regresión logística se elige por defecto, debido al caso de estudio, pero hay razones para esta decisión más allá de que fue el modelo usado en el estudio original: con respecto a esta tesis, la simpleza y popularidad de este modelo con respecto a otras opciones, permite ilustrar de una manera más práctica los contrastes entre metodologías, que son el punto central de este escrito. Además, considerando las medidas de la discriminación, la interpretación de los modelos de regresión logística como probabilidades de Bernoulli, facilita obtener estas medidas y controlar factores al comparar grupos protegidos.

En cuanto a los detalles del modelo para cada metodología, son puntos que ya se han justificado y que en el caso de la metodología bayesiana se seguirán revisando en la siguiente

sección. Para la metodología frecuentista, el uso de la verosimilitud logarítmica negativa (NLL) como costo y el método de mínimos cuadrados iterativamente ponderados (IWLS), son decisiones que también se toman del trabajo original, pero por lo general, es bastante común elegir ambas características en el aprendizaje computacional, y ello se puede ver en que son decisiones por defecto en las funciones que se usaron para entrenar este modelo, tanto en R (para caso del trabajo original) como en TensorFlow (en el caso de la reproducción implementada).

Por su parte, las decisiones para el caso bayesiano ya se expusieron antes. El costo se compone tanto de la NLL como de la Divergencia KL (entre la distribución a priori y la distribución variacional), pues se usa la inferencia variacional, y este costo representa la diferencia entre la distribución variacional y la distribución a posteriori. En cuanto al optimizador Adam y el uso de mini-lotes, son decisiones que se justificaron en el ejemplo con datos sintéticos: Adam se seleccionó por su buen rendimiento promedio, y los mini-lotes porque son un compromiso adecuado entre el método por lote y el método estocástico. Finalmente, la distribución normal de campo medio simplemente representa que se asignan distribuciones normales estándar a los coeficientes; en este caso la media a priori se escribe como un vector de ceros, pues son las medias a priori de todos los coeficientes, y la varianza a priori se escribe como la matriz identidad, pues representa que los coeficientes tienen, a priori, varianzas unitarias, y que la distribución de campo medio es la factorización de las distribuciones individuales (por eso es la matriz identidad, los unos en la diagonal son estas varianzas unitarias, y los ceros en el resto de la matriz indican dicha factorización entre distribuciones individuales, como partes la distribución conjunta de campo medio).

Finalmente puede ilustrarse la obtención de resultados con cada modelo. Ya que esto será igual para reincidencia general y violenta, se presenta un único esquema (los conceptos ilustrados dependen sólo de los modelos, no de los datos usados):

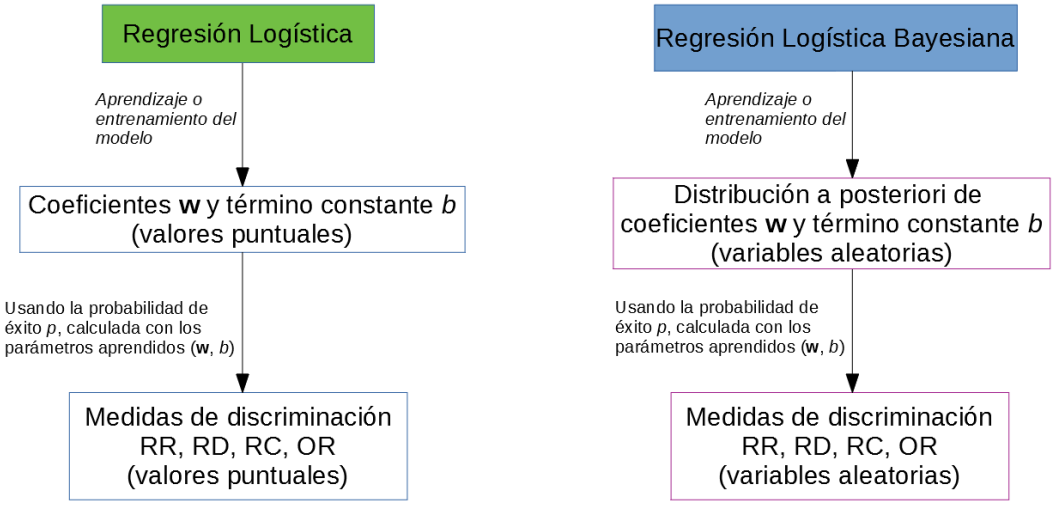


Figura 10.4.3. Esquema para la obtención de resultados.

Con este esquema se puede ver la diferencia fundamental entre los resultados para cada metodología, que ya se ha establecido en capítulos anteriores. Mientras en la regresión logística regular (frecuentista) se aprenden valores puntuales para los parámetros, en la regresión logística bayesiana se obtienen distribuciones a posteriori sobre los mismos. Esto se extiende también a las medidas de la discriminación que se obtienen con los parámetros aprendidos, que en el caso frecuentista tienen valores puntuales y, como se revisará a detalle en la siguiente sección, para el caso bayesiano son variables aleatorias, pues se obtienen en función de parámetros que son aleatorios en sí, ya que son descritos por las distribuciones a posteriori aprendidas.

10.3 IMPLEMENTACIÓN BAYESIANA

A continuación se implementará el mismo problema con los mismos datos, pero se empleará una regresión logística bayesiana mediante inferencia variacional, descrita con más detalle en capítulos anteriores. En esta sección el enfoque estará sobre los resultados obtenidos.

Como se expresó antes, en el caso bayesiano se establece una distribución de probabilidad a priori para los parámetros a estimar (en este caso los coeficientes de regresión logística), y mediante la inferencia variacional se obtiene una distribución a posteriori tras observar los datos. En este caso la distribución empleada es la llamada distribución normal de campo medio (*mean-field normal distribution*) (Keng, 2017). Sin ahondar en los detalles de esta distribución, básicamente es el equivalente a seleccionar una distribución normal para cada uno de los coeficientes de regresión logística, cuyos parámetros (media y varianza) se actualizan o se "aprenden" de los datos observados. TensorFlow Probability lidia con los detalles de esta formulación internamente y permite manipular la distribución sobre cada parámetro con facilidad.

Por defecto, la distribución a priori es la distribución normal estándar para cada coeficiente (es decir con media igual a cero y varianza unitaria)[69]. Ya que en la distribución a priori se suele incorporar el conocimiento previo de los problemas, si no se considera que haya tal conocimiento, suelen seleccionarse varianzas elevadas para representar que el valor verdadero de las cantidades estimadas podría estar en casi cualquier punto de la recta numérica. Por el momento se presentarán los resultados con esta varianza unitaria, y en el anexo de resultados adicionales se mostrará el efecto de seleccionar distintas opciones en la distribución a posteriori aprendida, donde se corrobora que a pesar de no tener alguna razón en particular para pensar que los coeficientes estarán cerca del cero, la varianza unitaria es suficiente para que tras aprender de los datos los coeficientes lleguen a un valor adecuado, y muy similar al que llegan cuando se usan grandes varianzas a priori[70]. Se presentarán primero los resultados para la reincidencia general, exponiendo varios detalles importantes del contraste entre el método bayesiano y frecuentista, los cuales se extienden al modelo para reincidencia violenta, pero será más sencillo y comprensible presentar estos detalles usando sólo la reincidencia general, y después presentar los resultados para la reincidencia violenta tras la discusión sobre las ventajas del método.

69 Es crucial resaltar que esta distribución normal sobre los coeficientes (que podría ser otro tipo de distribución) es lo que hace de este un método bayesiano, y no la distribución de Bernoulli considerada en cualquier regresión logística, bayesiana o no.

70 Uno de los atractivos de la metodología bayesiana es su regularización inherente: es común aplicar regularizadores a la regresión lineal, y demás métodos del aprendizaje computacional. Dichos regularizadores provocan que durante el aprendizaje no sólo se minimice la diferencia entre predicciones y salidas, que es el objetivo principal, sino también la magnitud de los "pesos" o coeficientes (Murphy, 2012, pp. 225-227). Usar una distribución a priori centrada en el cero es un método de regularización porque incorpora, en forma de conocimiento previo, la consideración de que los coeficientes o pesos deberían estar cerca del cero (Gelman et al., 2013, pp. 367-368), (*Prior distribution and regularization*, s. f.).

Coeficientes de Regresión Logística (componentes del vector **w** y valor de *b*)				
Atributo correspondiente en el vector **x**	*Distribuciones a priori*		*Distribuciones a posteriori*	
	Media	Varianza	Media	Varianza
Ninguno (el coeficiente es la constante *b*)	0.0	1.0	-1.489185	0.001074
Factor de género: femenino (*gender_factor*)	0.0	1.0	0.270534	0.007617
Factor de edad: mayor de 45 (*age_factor*)	0.0	1.0	-1.415583	0.010041
Factor de edad: menor de 25 (*age_factor*)	0.0	1.0	1.280887	0.002522
Factor de raza: afroamericano (*race_factor*)	0.0	1.0	0.448626	0.001319
Factor de raza: asiático (*race_factor*)	0.0	1.0	-0.346765	0.160633
Factor de raza: hispano (*race_factor*)	0.0	1.0	-0.464798	0.015907
Factor de raza: nativo americano (*race_factor*)	0.0	1.0	0.726923	0.460661
Factor de raza: otro (*race_factor*)	0.0	1.0	-0.799635	0.016925
Conteo de antecedentes (*priors_count*)	0.0	1.0	0.254689	0.000136
Factor de crimen: delito menor *(crime_factor)*	0.0	1.0	-0.321789	0.002532
Reincidencia en dos años (*two_year_recid*)	0.0	1.0	0.703154	0.001101

Tabla 10.7. Distribuciones de los coeficientes en la implementación bayesiana.

Al comparar con los resultados obtenidos con el método frecuentista, es evidente que los valores de los coeficientes (tomando como referencia las medias posteriori) tienen diferencias con los valores de la reproducción, pero esto se debe en parte a que el método bayesiano empleó un esquema de optimización distinto, que también podría usarse con un método frecuentista (esto se discute en una parte del anexo para resultados adicionales)[71]. Por ahora, para ilustrar lo que está sucediendo, se puede observar la distribución a priori, que es igual para todos los coeficientes, y las distribuciones a posteriori que resultan después del entrenamiento, es decir, después de aprender de los datos.

71 El hecho de que las diferencias numéricas entre resultados frecuentistas y bayesianos, se deben en parte al cambio en los esquemas de optimización, y no sólo a las diferencias conceptuales entre las dos metodologías, es un detalle persistente en los diversos resultados de este escrito.

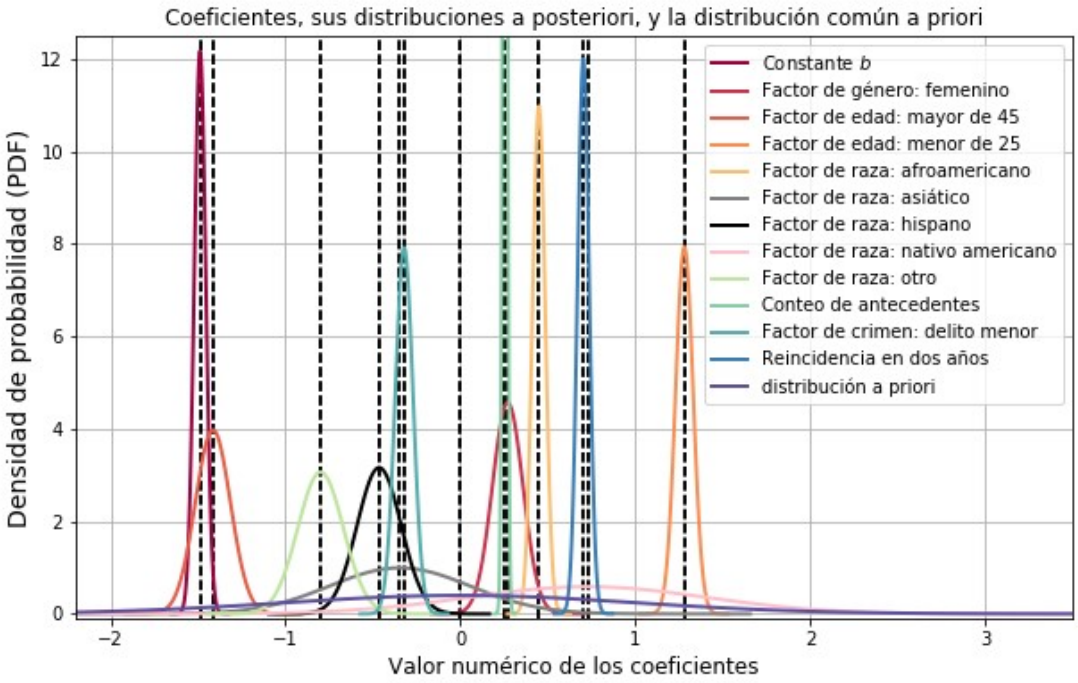


Figura 10.5.1. Distribuciones para los coeficientes de la regresión logística bayesiana. Las líneas punteadas son la media de cada una de las distribuciones representadas.

Como se puede apreciar en la Figura 10.5.1, y en las medias y varianzas de la Tabla 10.7, hay mayor incertidumbre en algunos coeficientes. Los coeficientes para los que hay mayor certeza después de observar los datos tienen distribuciones a posteriori con menor varianza (que por lo tanto son más angostas). Por ejemplo, las distribuciones a posteriori para la constante *b*, los coeficientes para raza afroamericana, reincidencia de dos años y conteo de antecedentes, muestran que la incertidumbre es baja, pues la varianza es pequeña (en particular, la varianza de la distribución del conteo de antecedentes es tan pequeña que casi parece una linea vertical). A su vez, las distribuciones para los coeficientes de factores de raza nativo-americano y asiático, tienen una varianza muy elevada, resultando en distribuciones muy anchas, que son difíciles de apreciar en la Figura 10.5.1. Para poder apreciarlas mejor, las distribuciones de estos dos factores se grafican por separado en la Figura 10.5.2. (nótese el cambio en la magnitud máxima del eje vertical):

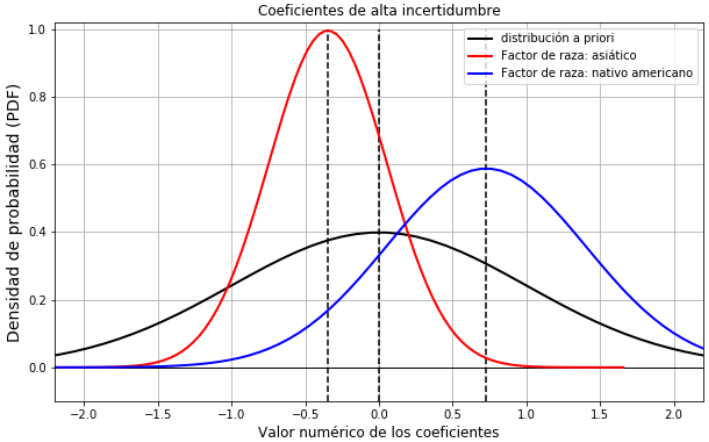


Figura. 10.5.2. Distribuciones a posteriori para los coeficientes de mayor incertidumbre (factores de razas asiática y nativo americana).

Como se puede observar, no hay una reducción tan significativa en la incertidumbre tras aprender de los datos para estos dos coeficientes. En ambos gráficos también se puede apreciar la distribución a priori, que tiene la varianza más elevada: todas las otras distribuciones eran como esta antes de aprender de los datos. Esta distribución se extiende a lo largo del eje horizontal y es "baja" en el eje vertical, pues la densidad de probabilidad está distribuida en un rango amplio de valores.

Una varianza baja para un coeficiente quiere decir que hay certeza en la relación que el modelo describe para la característica asociada a dicho coeficiente, mientras que una varianza elevada quiere decir que el modelo tiene incertidumbre, y ello se ve reflejado en distribuciones anchas, que ponen su densidad de probabilidad sobre un rango más amplio de posibles valores, como en el caso de los factores de alta incertidumbre de la figura 10.5.2.

Los resultados son congruentes. Ninguna distribución tiene varianza mayor a la varianza unitaria a priori, pues tras observar datos hay más certeza, nunca más incertidumbre[72]. Del mismo modo, los resultados tienen sentido considerando los datos observados: por ejemplo, en los datos hay 3175 elementos de raza negra, y tan sólo 31 de raza asiática y 11 de raza nativo americana, es por esto que hay una certeza considerable para el coeficiente asociado con la raza afroamericana, mientras que las distribuciones a posteriori de las otras dos razas tienen las varianzas más elevadas de entre todos los coeficientes, pues el modelo sólo pudo aprender de un número muy limitado de ejemplos.

Como se mencionó antes, y como se puede apreciar en las Figuras 10.5.1 y 10.5.2, la varianza unitaria a priori es adecuada, pues se asigna suficiente probabilidad a los valores encontrados de cada coeficiente, logrando probabilidades a posteriori adecuadas al terminar el entrenamiento. Es decir, la varianza a priori expresa suficiente incertidumbre para que los coeficientes converjan a los valores expuestos (una varianza muy pequeña no permitiría aprender valores fuera de la vecindad inmediata del cero). Para no desviar más la discusión, en el anexo para resultados adicionales se muestra empíricamente este punto, probando varianzas mucho mayores y mostrando que de cualquier modo se converge a valores muy próximos a los obtenidos con varianza unitaria.

Ventajas Generales del Método Bayesiano en la Evaluación del Sesgo

Ya se planteó el tipo de resultados que se obtienen al realizar la regresión logística bayesiana, pero antes de hablar de las medidas de discriminación, que son las cantidades centrales para este caso de estudio (en particular el riesgo relativo), se puede hablar de las ventajas en términos de los coeficientes de la regresión.

En el estudio original, al basarse en el método frecuentista, los coeficientes se presentan con su respectiva significación estadística. ProPublica no menciona directamente estos datos en sus conclusiones o análisis para los modelos de regresión logística, pero aún así se presentan con los resultados. A primera instancia podría parecer que la cuantificación de la incertidumbre que ofrecen las varianzas de las distribuciones a posteriori es algo que podría observarse en la significación estadística de los coeficientes, pero no es así.

Los niveles de significación posibles para los coeficientes son: 0, 0.001, 0.01, 0.05, 0.1 y 1. Cada nivel representa la probabilidad máxima de la hipótesis nula para el coeficiente en cuestión, es decir, la probabilidad de que el factor asociado con el coeficiente no tenga efecto alguno en la predicción del puntaje COMPAS dentro del modelo de regresión logística (o bien, que dicho

72 Por esto es importante elegir una distribución a priori adecuada, ya sea que tenga una varianza suficientemente elevada para representar la ignorancia con respecto a los valores, o que incorpore adecuadamente el conocimiento previo. Si bajo la distribución a priori, los datos observados son técnicamente imposibles, y la densidad de probabilidad que se les asigna antes de verlos es mínima, entonces muy difícilmente la probabilidad a posteriori será adecuada, pues la actualización se da acorde al teorema de Bayes, donde la probabilidad a posteriori de un parámetro está determinada por el producto de la probabilidad a priori del mismo y su verosimilitud acorde a los datos, entonces, a pesar de tener una verosimilitud elevada porque los datos lo indican, si su probabilidad a priori era mínima, el aprendizaje será lento y en muchas ocasiones intratable en el sentido práctico.

coeficiente sea cero). En los resultados de COMPAS, la mayoría de los coeficientes lograron la significación estadística de 0, que no se refiere a una probabilidad de cero para la hipótesis nula, pero sí menor a 0.0001 (*«p» Value of 0.000?*, s. f.). Sin embargo, los coeficientes para factor de género femenino, factor de raza asiática, y factor de raza nativo-americana, alcanzaron niveles de 0.001, 1, y 0.1, respectivamente.

Parecería intuitivo que para estos coeficientes con menor significación, se pudiera hablar de una mayor incertidumbre, y parece ser congruente el hecho de que los dos factores que tuvieron una mayor varianza a posteriori para sus coeficientes (y por lo tanto mayor incertidumbre) en el método bayesiano, sean los mismos que tienen menor significación estadística[73]. El problema con el método frecuentista es que la significación estadística no brinda información de utilidad en este contexto, sobre todo considerando que dicho método se basa en el concepto de experimentos repetibles.

La incompatibilidad entre la significación estadística y la cuantificación de la incertidumbre se puede observar fácilmente en dos situaciones presentes en los resultados: primeramente, aunque los dos coeficientes con mayor varianza a posteriori del método bayesiano son los mismos con menor significación estadística en el método frecuentista, el que tiene mayor varianza a posteriori de los dos (factor de raza nativo-americana) tiene también la mayor significación estadística de ambos. En segundo lugar, aunque el coeficiente para el factor de edad mayor de 45 tiene una mayor varianza a posteriori dentro del método bayesiano que aquel asociado con el factor de género femenino, logró tener la misma significación que los demás coeficientes, mayor a la del factor de género femenino.

Todo esto se debe a que la significación estadística no considera mas que la probabilidad de que un coeficiente sea cero, independientemente de la certeza asociada a su valor, así que si hubiera un coeficiente con un valor muy pequeño, pero del cual se tiene una gran certeza, de cualquier modo es probable que tenga una significación pequeña. De la misma manera, un coeficiente del cual se tiene una incertidumbre considerable, pero se espera que tenga un valor elevado, puede tener una significación alta.

Además, la significación se considera por niveles, y no suele tener sentido tratar de ver cuál tiene mayor significación de entre los que están en el mismo nivel (aunque técnicamente podría hacerse, con base en los valores p, que determinan la significación). Tampoco hay que perder de vista la noción de experimentos repetibles que se considera en el caso frecuentista: no es lo más coherente considerar los datos de este estudio como el producto de un experimento que se pueda reproducir. Finalmente, cabe mencionar que aún para casos que encajan con la noción de experimentos reproducibles, suele hallarse que los valores p que determinan la significación dependen del muestreo y no son reproducibles en sí. Además hay críticas considerables al uso generalizado de los niveles por defecto de significación estadística, que suelen considerarse para determinar si algún resultado tiene o no dicha significación (en especial este es el caso del nivel de 0.05), aún en contextos donde los niveles por defecto no son adecuados (Dahiru, 2008).[74]

Por todas estas razones, no se puede hablar de una cuantificación de la incertidumbre en el método frecuentista, mientras que las varianzas a posteriori del método bayesiano son precisamente eso. Una gran ventaja del método bayesiano, es que hay ocasiones en las que es necesario ofrecer predicciones o inferencias a pesar de la incertidumbre. En este caso, de todos los factores considerados para la regresión, ProPublica sólo menciona los coeficientes para raza negra, edad

73 Otro punto que pareciera sugerir esta interpretación, es que el coeficiente del factor de género femenino tiene una significación menor a los factores de raza afroamericana y edad menor a 25 años, pues en el anexo para resultados adicionales se puede observar que el riesgo relativo para elementos de género femenino cambia al considerar una modificación propuesta por (W. Flores et al., 2016) en la estructura del modelo, al grado de contradecir una conclusión secundaria en el análisis de ProPublica (*propublica/compas-analysis*, 2016/2020), (Larson et al., s. f.), lo cual no ocurre para elementos considerando los otros dos factores. Para más detalles al respecto, véase la sección correspondiente en el anexo.

74 En (Rozeboom, 1960) se presenta una crítica general de la significación estadística con una discusión más profunda.

menor a 25 años, y género femenino para el análisis de la discriminación (encima de esto, el artículo original está centrado totalmente en el factor de raza negra). Sin embargo, considerando la posibilidad de obtener medidas de discriminación con estos modelos para los factores de razas asiática y nativo americana, los resultados frecuentistas simplemente pueden decir que al no haber significación estadística suficiente, las medidas obtenidas no serán válidas, mientras que los resultados bayesianos pueden ofrecer respuestas e incorporar la incertidumbre en las mismas de una forma explícita y estructurada.

Otro componente valioso del método bayesiano que ya se ha mencionado antes, es que esta misma incertidumbre es una guía para la recolección de datos futuros, pues para el caso de este modelo, denota que sería bueno juntar más datos que involucren elementos de razas asiática o nativo americana. Esto último es una gran desventaja del método frecuentista en este contexto, pues la significación estadística no es una buena guía, ya que coeficientes considerablemente certeros pueden tener baja significación sólo por tener valores cercanos al cero, y coeficientes para los cuales hay considerable incertidumbre, pueden alcanzar la significación estadística si su rango plausible, a pesar de ser muy amplio, está suficientemente alejado del cero.

Todas estas ventajas son importantes en aplicaciones más allá del estudio de prejuicios o sesgos en herramientas como COMPAS, pero es importante enfatizar lo que ofrecen en este contexto y en otros relacionados con este tema. Al evaluar una herramienta, como COMPAS en este caso, el método bayesiano permite dar respuestas más informativas (esto se podrá ver más a fondo cuando se obtengan las medidas de discriminación más adelante), y estas respuestas no sólo sirven para dar una evaluación más robusta, considerando explícitamente la incertidumbre de las medidas obtenidas, sino que también son respuestas más intuitivas y útiles en general, comparadas a las respuestas escalares y la medida de significación estadística del método frecuentista. Además, la incertidumbre es indicativa de qué tipo de datos se necesita recolectar o explorar para dar una evaluación más certera. Todo esto sin mencionar que en casos como este, no parece razonable la noción de experimentos repetibles sobre la cual se basa el método ortodoxo. Así como el método bayesiano ofrece estas ventajas al evaluar una herramienta como COMPAS, otra temática interesante es observar estas mismas ventajas en el contexto de las herramientas en sí, es decir, usar clasificadores bayesianos para las tareas que dichas herramientas llevan a cabo (en el caso de COMPAS, esto podría ser equivalente a ofrecer una distribución de probabilidad sobre los puntajes de cada individuo, en vez de un número concreto del 1 al 10). De hecho, ya se mencionó que la mayoría de los trabajos relacionados se enfocan en tareas de este tipo. Sin embargo, este trabajo está basado en este caso de estudio, y como tal, se enfocará sólo al proceso de evaluación de la herramienta (descubrimiento de la discriminación). Aún así, es importante entender que estas ventajas pueden trasladarse fácilmente a otros procesos relacionados con el sesgo en el aprendizaje computacional, y con la automatización de decisiones en general.

Medidas de Discriminación con el Método Bayesiano

Las conclusiones de la evaluación de ProPublica que conciernen a los modelos de regresión logística se basan en la medida de riesgo relativo (RR), pero como en el caso de la implementación frecuentista, se pueden obtener las demás medidas pertinentes. En el caso frecuentista, estas medidas tienen un valor único, pues están basadas en los coeficientes de regresión logística, que en ese contexto son un valor puntual; sin embargo, para el caso bayesiano estos coeficientes tienen distribuciones de probabilidad.

Ya que ahora las medidas de discriminación bayesianas están basadas en variables aleatorias y no en escalares, hay varias maneras de proceder. Una opción es obtener un número de muestras para estas medidas, muestreando primero los coeficientes con base en sus distribuciones y usando las distintas muestras para calcular varios valores de la medida en cuestión, el resultado será entonces una serie de valores posibles para la medida, de la cual se puede obtener la media, el rango y la varianza. Otro método es simplemente derivar los rangos con base en las distribuciones de los

coeficientes que se usarán para calcular cada medida y determinar algún rango para los mismos con base en sus distribuciones (por ejemplo, considerando los valores tres desviaciones estándar por debajo y por encima de la media de cada coeficiente como sus valores máximo y mínimo), para posteriormente obtener el rango máximo y mínimo de la medida discriminatoria en cuestión.

En este caso se muestrearán las medidas por dos razones principales: en primer lugar, con este método se puede obtener una media para la medida discriminatoria, así como rangos basados en las muestras obtenidas. Y en segundo lugar, porque no hay razones para pensar que determinar un rango arbitrario para los valores de los coeficientes de antemano sea preferible en esta aplicación, sin mencionar que el proceso de muestreo es, en concepto, más simple[75]. Además, obtener muestras es computacionalmente sencillo (así que no hay beneficios prácticos significativos como para evitar obtener las muestras).

	RR		**RD**		**RC**		**OR**	
	Media	*Varianza*	*Media*	*Varianza*	*Media*	*Varianza*	*Media*	*Varianza*
Raza negra	1.418876	0.001453	0.077086	5.1e-05	0.905509	7.9e-05	1.567475	0.003228
Género Femenino	1.241396	0.007025	0.044426	0.000239	0.945544	0.00036	1.315198	0.013306
Edad menor a 25 años	2.435149	0.005034	0.264063	0.000165	0.67633	0.000266	3.604618	0.032974

Tabla 10.8.1. Medidas de la discriminación obtenidas por muestreo con el método bayesiano.

En la Tabla 10.8.2 se muestran los rangos en términos de límites inferiores y superiores. Es común establecer rangos usando dos o tres desviaciones estándar por arriba y debajo de la media para cantidades bajo una distribución normal, porque incluyen la mayoría de los valores posibles, pues casi toda la probabilidad está distribuida entre los puntos dentro de estos rangos y los valores fuera de los mismos son muy inusuales. Esta práctica se usa aunque la distribución se obtenga con base en muestras, ya que aunque se podría establecer un rango total usando el mínimo y máximo exactos de la muestra, éstos pueden ser valores extremos, que no son útiles para ilustrar los datos que uno puede observar usualmente si provienen de la misma distribución.

De las cuatro medidas, sólo la diferencia de riesgo (RD) puede tomar valores negativos, y además se limita a valores entre -1 y 1. Es evidente que estas medidas no tienen una distribución normal, por lo que no es adecuado establecer los rangos mencionados basados en desviaciones estándar, pues se basan en las características de la distribución normal. Una de las grandes ventajas del método de muestreo es que los rangos para cantidades secundarias como éstas, se pueden encontrar fácilmente usando percentiles, sin necesidad de analizar la distribución de las mismas y sus características. En este caso se usarán los percentiles correspondientes a 0.15% y 99.85%, pues aunque el rango basado en las tres desviaciones estándar (por encima y debajo de la media) se usa por lo práctico de tratar con desviaciones estándar al usar distribuciones normales, en realidad sólo es un rango que cubre aproximadamente el 99.7% de los valores bajo esa densidad de probabilidad (PDF) («68–95–99.7 Rule», 2020). Se usarán estos percentiles para tener consistencia con este concepto, cubriendo ese mismo porcentaje. El procedimiento es simple de implementar usando TensorFlow y NumPy.

75 Si se quisieran obtener datos como la media, la varianza, o el mismo rango de las medidas de discriminación, sin usar muestreo, sería necesario obtener una solución analítica basada en las distribuciones de los coeficientes empleados en la obtención de dicha medida. Esto puede hacerse porque ya se tienen las distribuciones de los coeficientes, pero no siempre es una tarea sencilla, y además es irrelevante y poco práctico para este trabajo.

	RR		RD		RC		OR	
	Mínimo	*Máximo*	*Mínimo*	*Máximo*	*Mínimo*	*Máximo*	*Mínimo*	*Máximo*
Raza negra	1.310733	1.535421	0.056951	0.099097	0.878097	0.930601	1.411009	1.745681
Género Femenino	1.006979	1.503201	0.001261	0.092903	0.885703	0.998447	1.00853	1.696606
Edad menor a 25 años	2.227872	2.650279	0.225956	0.302681	0.626601	0.724594	3.094453	4.182237

Tabla 10.8.2. Rangos para medidas de la discriminación con el método bayesiano.

Conclusiones con el Método Bayesiano y sus Ventajas

ProPublica sólo usa la medida de riesgo relativo (RR) en su estudio, pero hay ventajas del método bayesiano que se reflejan al obtener las otras medidas. Para empezar, con base en las varianzas, es evidente que hay menos incertidumbre con respecto a las medidas de diferencia de riesgo (RD) y tasa de selección (RC), que con respecto al RR y el radio de posibilidades (OR). Esto es interesante, ya que las cuatro medidas pueden usarse para cuantificar el mismo fenómeno, y como se mencionó en el capítulo correspondiente al análisis de la discriminación, distintas autoridades usan distintas medidas. Como en realidad sólo se usó el RR en este caso, queda fuera del enfoque de este trabajo determinar si sería mejor usar las medidas de RD o RC, que reflejan mayor certeza, pues una conclusión significativa requeriría un análisis más a fondo y varias pruebas complementarias[76]. Aún así, es notable cómo la estructura del método bayesiano facilita estudiar estas posibilidades.

Ahora bien, como ya se mencionó en la reproducción del modelo original, con base en el RR ProPublica concluye que los elementos de raza negra, género femenino, y edad menor a 25 años, tienen probabilidades 45%, 19.4%, y 2.5 veces mayores, respectivamente, de ser clasificados con un puntaje alto por COMPAS[77]. Estas respuestas cerradas y exactas son propias de un método frecuentista, donde al mostrar la significación estadística para los coeficientes, entonces son usados para calcular otras cantidades que ya no son sujetas a pruebas de significación estadística. En contraste, en el método bayesiano la incertidumbre de los coeficientes determina también la incertidumbre sobre estas cantidades derivadas de los mismos, pudiendo dar así una respuesta más elaborada al problema: de acuerdo al modelo, los elementos de raza negra tienen una probabilidad de entre 31.1% y 53.4% mayor de ser clasificados con un puntaje alto por COMPAS, que es, en promedio, 41.9% mayor. Los elementos de género femenino a su vez tienen una probabilidad entre 0.7% menor y 50.3% mayor de recibir un puntaje alto por COMPAS, siendo en promedio una probabilidad 24.1% mayor. Finalmente, los elementos menores de 25 años tienen una probabilidad entre 2.23 y 2.65 veces mayor de recibir puntajes altos, teniendo en promedio una probabilidad 2.44 veces mayor.

Estas respuestas son más sustanciosas, y los resultados numéricos tienen un significado útil e intuitivo, pues como se puede notar en la discusión anterior sobre la signficación estadística, aún si se buscaran conclusiones más completas incorporando la significación de cada coeficiente, la cuestión es que no tiene un significado que se pueda trasladar de forma sencilla para ser de utilidad en este contexto.

Otro elemento ventajoso es la transparencia de las suposiciones bayesianas. Las distribuciones a priori indican de manera formal tanto la naturaleza que se asume para las cantidades a deducir, como el conocimiento previo o suposiciones que se incorporan al proceso. En

76 Otra consideración es que el OR se calcula usando el RR y el RC. También podría ser interesante explorar cómo la incertidumbre de las medidas en las cuales se basa afectan su propia incertidumbre, y en general qué significa todo esto para su aplicación como una medida de la discriminación.

77 Aquí se obvia la explicación de que estas diferencias están basadas en casos donde se comparan elementos de raza negra con elementos de raza blanca que tienen atributos iguales a excepción de la raza, elementos de género femenino con elementos de género masculino con atributos iguales a excepción del género, etcétera.

el caso frecuentista, también hay suposiciones, pero no son explícitas y modificables como en el caso bayesiano. Entre dichas suposiciones frecuentistas sobresalen el concepto de que el fenómeno es repetible, sobre el cual se basa la significación estadística, y los rangos y valores aceptables de la significación misma, que son estándares insensibles al problema en particular (Cleophas et al., 2006), (Dahiru, 2008). Una forma práctica de ilustrar esto es con las mismas distribuciones a priori: así como se obtienen medidas de discriminación muestreando las distribuciones a posteriori, se puede hacer lo mismo con las distribuciones a priori. Las medidas de la discriminación obtenidas de esta forma, permiten ilustrar las suposiciones previas con respecto al fenómeno y a la misma herramienta COMPAS:

	RR		**RD**		**RC**		**OR**	
	Media	*Varianza*	*Media*	*Varianza*	*Media*	*Varianza*	*Media*	*Varianza*
Raza negra	1.033708	0.272585	-0.000433	0.035823	1.036578	0.272647	1.638217	4.518625
Género Femenino	1.032036	0.274539	-0.001079	0.035568	1.03673	0.274021	1.641158	4.786787
Edad menor a 25 años	1.034032	0.27076	3.5e-05	0.036082	1.035975	0.277092	1.641162	4.439555

Tabla 10.9.1. Medidas de la discriminación obtenidas por muestreo con el método bayesiano, usando las distribuciones a priori de los coeficientes.

Con estas medidas a priori, se puede observar la información que confiere el modelo antes del aprendizaje, más allá de que la distribución a priori para los coeficientes indica que se espera que los mismos tengan valores situados alrededor del cero. Estas medidas a priori permiten ver las suposiciones relevantes para la aplicación en turno, y dan significado al conocimiento previo expresado en términos de distribuciones a priori para coeficientes de regresión logística.

De estas medidas a priori, se puede establecer la posición del modelo antes de observar los datos, pero en términos del análisis de la discriminación, y no sólo del modelo de regresión logística. En la medida fundamental para el estudio, el RR, y también en el RD y la RC, se puede notar la imparcialidad a priori que asume el modelo, ya que un valor unitario para el RR y el RC, y un valor de cero para la RD, denotan igualdad en las probabilidades de recibir o no un beneficio, es decir, estos valores indican que no hay discriminación acorde a estas medidas; también en el caso del OR, un valor unitario representa imparcialidad, y considerando que las medias a priori son todas cero para los coeficientes, el OR debería tener, en teoría, este valor unitario, pero como se puede observar, la incertidumbre implícita en las distribuciones resulta en una varianza muy grande a priori para el OR, así que aunque el valor medio indica que en la mayoría de los casos se espera que haya un sesgo desfavorable para los grupos protegidos en cuestión (raza afroamericana, género femenino, edad menor a 25 años), la elevada incertidumbre para el OR expresa que el valor medio obtenido no está asociado con un conocimiento previo rígido y certero.

Los rangos a priori de estas medidas contienen más información que vale la pena analizar:

	RR		**RD**		**RC**		**OR**	
	Mínimo	*Máximo*	*Mínimo*	*Máximo*	*Mínimo*	*Máximo*	*Mínimo*	*Máximo*
Raza negra	0.083845	4.129251	-0.547096	0.54528	0.091181	4.024211	0.0501	19.007527
Género Femenino	0.086285	4.102349	-0.549655	0.548956	0.081914	4.078124	0.051349	20.671874
Edad menor a 25 años	0.085952	4.099819	-0.553762	0.547414	0.095386	4.148448	0.052561	18.266899

Tabla 10.9.2. Rangos para medidas de la discriminación a priori con el método bayesiano.

Nótese que para todas las medidas, sus medias, varianzas y rangos a priori son similares entre los tres factores. Primero se analizarán el RR y la RC, que tienen valores similares para media, varianza, mínimo y máximo. Recordando del capítulo de análisis de la discriminación, RR se refiere al radio de probabilidades de no recibir el beneficio y RC al radio de probabilidades de recibirlo. En este contexto no recibir el beneficio equivale a tener un puntaje COMPAS alto. El hecho de que el rango de ambas medidas esté aproximadamente entre 0.08 y 4, y que la media sea aproximadamente unitaria, quiere decir que el modelo asume a priori que no hay razón para creer que hay disparidades en la probabilidad de recibir un puntaje alto, pero a la vez establece que una persona de raza afroamericana, género femenino o edad menor a 25 años, podría tener una probabilidad entre 12 veces menor y 4 veces mayor (en términos de porcentaje esto es 92% menor y 300% mayor, respectivamente) que otro elemento con características equivalentes, a excepción de la raza, el género, o la edad.

En cuanto a la RD, su media cercana al cero indica que no hay razón para esperar discriminación a favor de ningún grupo entre individuos similares[78] que se distingan entre sí con base en su raza, género, o edad. El rango está aproximadamente entre -0.54 y 0.54, indicando que se espera una diferencia *absoluta*[79] del 54% como máximo entre las probabilidades de recibir un puntaje alto, y ésta puede ser a favor o en contra de los grupos representados por cada factor.

Por su parte el OR tiene una media de aproximadamente 1.64. Esto quiere decir que, estrictamente hablando, se asume a priori que poseer alguno de los factores, resulta en posibilidades 64% más grandes de recibir un puntaje alto[80]. Sin embargo, los rangos del OR se encuentran más o menos entre 0.05 y 20, es decir, a pesar de la aparente suposición de prejuicio dada por la media, los rangos indican que en realidad se espera tener posibilidades cerca de 20 veces mayores para elementos que poseen alguno de estos factores, pero que también podrían ser 20 veces menores, reflejando la elevada varianza a priori. En realidad la distinción entre factores en este caso sólo tiene el propósito de ilustrar el proceso: todos los coeficientes tienen la misma distribución a priori, y por lo tanto, el muestreo a priori de las medidas de discriminación para cualquier factor se puede interpretar como el de las medidas a priori para cualquier otro factor[81].

Como se ha visto hasta el momento en la implementación bayesiana, esta metodología ofrece información de gran utilidad que no está presente al usar un método frecuentista. La capacidad de usar esta información no se limita a los parámetros del modelo clasificador, sino que se extiende para las medidas de la discriminación basadas en dichos parámetros. Esta información es fácil de interpretar con las distribuciones y estadísticas, y no sólo extiende la perspectiva para obtener conclusiones de los resultados, sino también para determinar qué se está asumiendo con el modelo antes de obtener dichos resultados. Sin embargo, como ya se mencionó antes, hay un compromiso o *trade off* inherente al método bayesiano, pues es evidente que requiere más trabajo en la elaboración de modelos y en la obtención de resultados, que a pesar de ser sencillos conceptualmente, implican llevar acabo diversas tareas para su obtención, inexistentes en el caso frecuentista. Estas ventajas y desventajas en el nivel abstracto, que se refiere principalmente a la

78 Es preciso reiterar que cuando se habla de individuos similares distinguidos por alguno de los factores, se habla de aquellos que tienen todas las características por defecto, a excepción del factor en cuestión. Dichas características por defecto están determinadas por el modelo de regresión, y ya se detallaron en la sección previa donde se reprodujo el trabajo original.

79 La diferencia absoluta de probabilidad que se da con la RD es muy diferente a la diferencia dada por el RR mencionada a lo largo del escrito y en el mismo estudio de ProPublica. Por ejemplo, si las probabilidades son 0.15 y 0.1, la primera probabilidad es 50% mayor a la primera, pero su diferencia de probabilidad en el contexto de la RD, es decir, la diferencia absoluta, es de tan sólo 5%.

80 Es importante recordar que el término "posibilidades" se refiere al término en inglés *odds* que puede tener diversas traducciones (Tapia-Granados, 1997). Las posibilidades de un elemento en este contexto son la probabilidad de recibir puntaje alto entre la probabilidad de no recibirlo. El OR indica cuán mayor es ese radio para un grupo comparado a otro.

81 Puesto de otra forma, las diferencias entre factores en las medidas de la discriminación a priori, sólo se observan por la aleatoriedad del proceso de muestreo.

esencia de las dos metodologías, es el enfoque principal de este trabajo, y es el tipo de *trade-off* que se pretende investigar.

Aunque no se desarrolle en el trabajo original de ProPublica, también se pueden obtener resultados para el rendimiento de los modelos en términos de su generalización y precisión, permitiendo compararlos también en ese sentido: esto se estudia en el anexo de resultados adicionales. Aunque ProPublica no lo menciona, los resultados que presentan para los modelos de reincidencia general y violenta, incluyen el criterio AIC, que determina si el ajuste del modelo es adecuado para los datos (*Akaike Information Criterion - an overview | ScienceDirect Topics*, s. f.), y en particular, si es superior al ajuste de un modelo que sólo use la constante *b* y ningún otro atributo. Esto no está en el enfoque del trabajo, por lo que se omite en este capítulo (como en el trabajo original), pero se ve con más detalle en el anexo, donde también se muestra que todos los modelos frecuentistas y bayesianos de este capítulo fueron adecuados acorde al criterio.

Reincidencia Violenta

En esta sección se muestran los resultados para los coeficientes y las medidas de la discriminación en el caso de reincidencia violenta. Los detalles de cada punto ya fueron discutidos para la reincidencia general, así que en esta parte sólo se presentan directamente los resultados y las conclusiones que se pueden derivar de ellos con respecto al caso de estudio: no se repetirán las diversas discusiones que ya se desarrollaron para la reincidencia general.

Coeficientes de Regresión Logística (componentes del vector **w** y valor de *b*)				
Atributo correspondiente en el vector **x**	*Distribuciones a priori*		*Distribuciones a posteriori*	
	Media	Varianza	Media	Varianza
Ninguno (el coeficiente es la constante *b*)	0.0	1.0	-2.085440	0.003090
Factor de género: femenino (*gender_factor*)	0.0	1.0	-0.704297	0.009678
Factor de edad: mayor de 45 (*age_factor*)	0.0	1.0	-1.763843	0.025178
Factor de edad: menor de 25 (*age_factor*)	0.0	1.0	3.046187	0.005902
Factor de raza: afroamericano (*race_factor*)	0.0	1.0	0.597401	0.006563
Factor de raza: asiático (*race_factor*)	0.0	1.0	-0.679483	0.335395
Factor de raza: hispano (*race_factor*)	0.0	1.0	-0.247318	0.016879
Factor de raza: nativo americano (*race_factor*)	0.0	1.0	-0.008983	0.535349
Factor de raza: otro (*race_factor*)	0.0	1.0	-0.344551	0.092972
Conteo de antecedentes (*priors_count*)	0.0	1.0	0.126147	0.000145
Factor de crimen: delito menor *(crime_factor)*	0.0	1.0	-0.131252	0.009287
Reincidencia en dos años (*two_year_recid*)	0.0	1.0	0.926821	0.007675

Tabla 10.10. Distribuciones de los coeficientes en la implementación bayesiana (reincidencia violenta)

Como en el caso de reincidencia general, los coeficientes difieren de los obtenidos con el método frecuentista, aunque en ambos casos las disparidades no son extremas. A continuación se mostrarán representaciones gráficas para las distribuciones a posteriori de los coeficientes, como se hizo en la sección anterior. La distribución a priori para todos los coeficientes sigue siendo la distribución normal estándar.

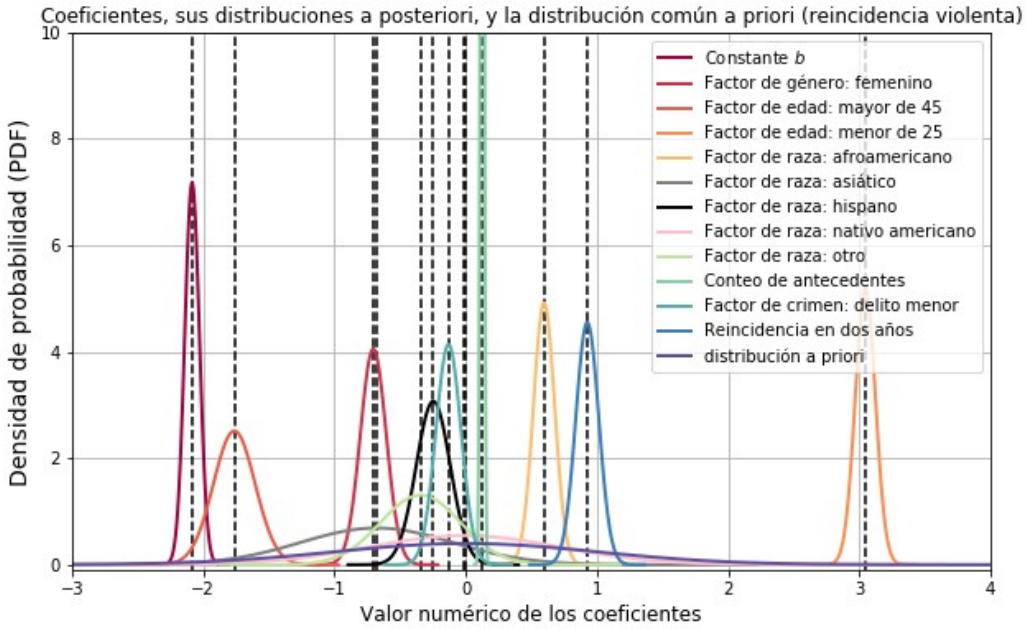


Figura. 10.6.1. Distribuciones para los coeficientes de la regresión logística bayesiana para los datos de reincidencia violenta. Las líneas punteadas son la media de cada una de las distribuciones representadas.

Como podría esperarse de la exploración de datos que se llevó acabó en la reproducción de la implementación frecuentista, nuevamente los factores de razas asiática y nativo americana, tienen varianzas elevadas, y por lo tanto hay más incertidumbre en sus valores; esto se podía esperar porque en dicha exploración se puede ver que los elementos de estos dos grupos son limitados en la población. Nuevamente el conteo de antecedentes tiene una varianza muy pequeña, y toda su probabilidad se concentra en un rango reducido de valores. En general, la incertidumbre para cada coeficiente no es muy diferente de la observada en el caso de reincidencia general.

Como se hizo en el caso anterior, se graficarán por separado las distribuciones para los dos coeficientes de mayor incertidumbre, con la distribucion a priori, y en este caso también se añadirá la distribución del factor de edad menor a 25 años, pues como su media está cerca del extremo derecho en la Figura 10.6.1, no se aprecia tan bien como las demás distribuciones.

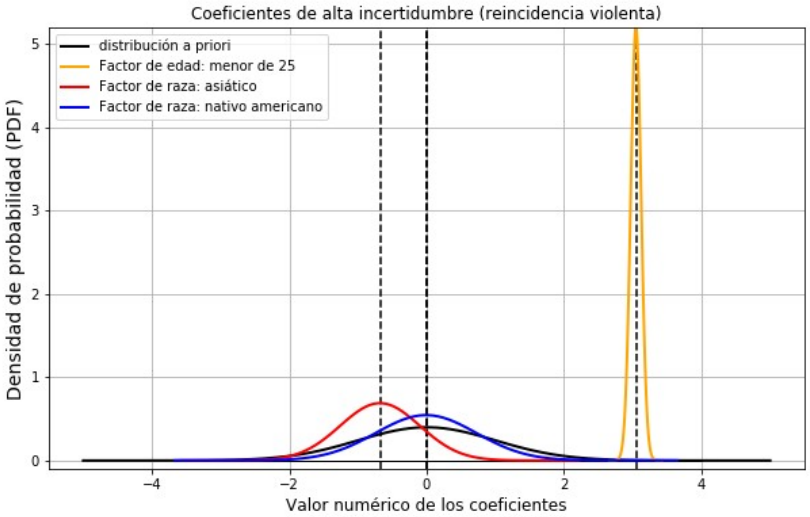


Figura. 10.6.2. Distribuciones a posteriori para los coeficientes de mayor incertidumbre (factores de razas asiática y nativo americana) y para el factor de edad menor a 25 años, en el caso de reincidencia violenta.

Similar al caso para reincidencia general, los factores de alta varianza o incertidumbre no son muy diferentes de la distribución a priori, y su densidad de probabilidad se encuentra sobre un rango amplio de valores. Esto es por los pocos datos disponibles, habiendo sólo 26 elementos de raza asiática, y 7 de raza nativo americana. Nuevamente se observa que la varianza a priori es adecuada, pues la distribución a priori asigna suficiente densidad de probabilidad a las medias aprendidas. Esto ya se estudió a fondo en el anexo para el caso de reincidencia general, así que en este caso ya no se ahonda en el tema.

	RR		**RD**		**RC**		**OR**	
	Media	*Varianza*	*Media*	*Varianza*	*Media*	*Varianza*	*Media*	*Varianza*
Raza negra	1.670232	0.012123	0.074122	0.000156	0.916633	0.000201	1.824335	0.02175
Género Femenino	0.52556	0.002365	-0.052487	3.5e-05	1.059036	4.6e-05	0.496541	0.002382
Edad menor a 25 años	6.545428	0.069367	0.612367	0.000268	0.311384	0.000408	21.110197	2.649327

Tabla 10.11.1. Medidas de la discriminación obtenidas por muestreo con el método bayesiano (reincidencia violenta).

	RR		**RD**		**RC**		**OR**	
	Mínimo	*Máximo*	*Mínimo*	*Máximo*	*Mínimo*	*Máximo*	*Mínimo*	*Máximo*
Raza negra	1.368914	2.020165	0.040099	0.114055	0.871194	0.955041	1.433888	2.314531
Género Femenino	0.398772	0.68496	-0.069225	-0.034332	1.03835	1.078748	0.371295	0.659402
Edad menor a 25 años	5.803226	7.37395	0.561498	0.657492	0.254908	0.374211	16.79482	26.400984

Tabla 10.11.2. Rangos para medidas de la discriminación con el método bayesiano (reincidencia violenta).

Con la metodología descrita para la reincidencia general, se obtienen las medidas de la discriminación con el método bayesiano para la reincidencia violenta. En el trabajo original de ProPublica, con base en los valores hallados para el riesgo relativo (RR), se concluyó que para el caso de reincidencia violenta los elementos de raza negra tienen una probabilidad 77.3% mayor de recibir puntajes altos, mientras que los elementos menores a 25 años tienen una probabilidad 7.4 veces mayor. Siguiendo la misma lógica se podría concluir que los elementos de género femenino tienen una probabilidad 49.2% menor de recibir puntajes altos, pero esto no se hizo en el trabajo original de ProPublica.

Nuevamente las medidas de diferencia de riesgo (RD) y tasa de selección (RC), son las que presentan menor varianza, y por lo tanto menor incertidumbre. Independientemente de esto, las conclusiones se basan en el RR, y con el método bayesiano se puede concluir que: los elementos de raza negra tienen una probabilidad entre 36.9% y 102% mayor de recibir un puntaje alto para reincidencia violenta, siendo en promedio 67% mayor. Los elementos de género femenino tienen una probabilidad entre 60.1% y 31.5% menor de recibir puntajes altos para reincidencia violenta, y es 52.6% menor en promedio. Finalmente, los elementos que tienen menos de 25 años tienen una probabilidad entre 5.8 y 7.37 veces mayor de recibir puntajes altos, que es 6.55 veces mayor en promedio.

Las conclusiones a fin de cuentas no contradicen a las del caso frecuentista, pues el incremento de probabilidad de obtener puntajes altos en reincidencia violenta persiste para elementos de raza negra y menores a 25 años, sin mayores desviaciones entre el valor bayesiano medio y el valor frecuentista. Lo mismo puede decirse para la disminución de probabilidad para elementos de género femenino. Sin embargo, nuevamente las respuestas bayesianas son más completas y ofrecen información con respecto a la incertidumbre en los valores usados para estas conclusiones.

En este caso no tiene sentido revisar las suposiciones a priori del modelo con respecto a las medidas de discriminación, pues los resultados serán similares al caso de reincidencia general: todos los coeficientes en ese caso tenían la misma distribución normal estándar a priori, por lo que se llegará a resultados equivalentes a los expuestos en esa sección[82].

REFERENCIAS

68–95–99.7 rule. (2020). En *Wikipedia*. https://en.wikipedia.org/w/index.php?title=68%E2%80%9395%E2%80%9399.7_rule&oldid=963007052

Akaike Information Criterion—An overview | ScienceDirect Topics. (s. f.). Recuperado 28 de junio de 2020, de https://www.sciencedirect.com/topics/medicine-and-dentistry/akaike-information-criterion

Cleophas, T. J., Zwinderman, A. H., & Cleophas, T. F. (2006). The Interpretation of the P-Values. En *Statistics Applied to Clinical Trials* (pp. 103–115). Springer Netherlands. https://doi.org/10.1007/978-1-4020-4650-6_9

82 Esto no sólo aplica a los coeficientes de reincidencia violenta en este caso, sino que se extiende a cualquier otro modelo donde los coeficientes y el término constante b tengan la distribución normal estándar como su distribución a priori. Como se comentará en el próximo capítulo, estos mismos resultados se extienden a ese caso.

Dahiru, T. (2008). P – VALUE, A TRUE TEST OF STATISTICAL SIGNIFICANCE? A CAUTIONARY NOTE. *Annals of Ibadan Postgraduate Medicine*, *6*(1), 21-26.

Documentation for rpy2—Rpy2 3.3.1 documentation. (s. f.). Recuperado 15 de junio de 2020, de https://rpy2.github.io/doc/latest/html/index.html

Dutang, C. (2017). *Some explanations about the IWLS algorithm to fit generalized linear models*. https://hal.archives-ouvertes.fr/hal-01577698

equivant. (2019, abril 4). Practitioner's Guide to COMPAS Core. *Equivant*. https://www.equivant.com/practitioners-guide-to-compas-core/

Gelman, A., Stern, H. S., Carlin, J. B., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. Chapman and Hall/CRC.

Glm.fit function | R Documentation. (s. f.). Recuperado 15 de junio de 2020, de https://www.rdocumentation.org/packages/scidb/versions/1.2-0/topics/glm.fit

Keng, B. (2017, abril 3). *Variational Bayes and The Mean-Field Approximation*. Bounded Rationality. http://bjlkeng.github.io/posts/variational-bayes-and-the-mean-field-approximation/

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (s. f.). *How We Analyzed the COMPAS Recidivism Algorithm*. ProPublica. Recuperado 2 de mayo de 2020, de https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

MLPR w3a—Machine Learning and Pattern Recognition. (s. f.). Recuperado 24 de mayo de 2020, de https://www.inf.ed.ac.uk/teaching/courses/mlpr/2018/notes/w3a_intro_classification.html

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT press.

«p» value of 0.000? (s. f.). ResearchGate. Recuperado 20 de junio de 2020, de https://www.researchgate.net/post/p_value_of_0000

Prior distribution and regularization. (s. f.). Recuperado 16 de junio de 2020, de https://compbio.soe.ucsc.edu/html_format_papers/hughkrogh96/node6.html

Propublica/compas-analysis. (2020). [Jupyter Notebook]. ProPublica. https://github.com/propublica/compas-analysis (Original work published 2016)

R: Fitting Generalized Linear Models. (s. f.). Recuperado 15 de junio de 2020, de https://stat.ethz.ch/R-manual/R-patched/library/stats/html/glm.html

Romei, A., & Ruggieri, S. (2014). A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review*, *29*(5), 582–638.

Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin*, *57*, 416-428. https://doi.org/10.1037/h0042040

Tapia-Granados, J. A. (1997). Posibilidades, oportunidades, momios: Un comentario sobre la traducción del término odds. *Salud Pública de México*, *39*(1), 69-71.

Tensorflow/probability. (s. f.). GitHub. Recuperado 15 de junio de 2020, de https://github.com/tensorflow/probability

Tfp.glm.fit | TensorFlow Probability. (s. f.). TensorFlow. Recuperado 15 de junio de 2020, de https://www.tensorflow.org/probability/api_docs/python/tfp/glm/fit

Violent Crime. (s. f.). FBI. Recuperado 29 de abril de 2020, de https://ucr.fbi.gov/crime-in-the-u.s/2010/crime-in-the-u.s.-2010/violent-crime/violent-crime

W. Flores, A., Bechtel, K., & Lowenkamp, C. (2016). False Positives, False Negatives, and False Analyses: A Rejoinder to “Machine Bias: There’s Software Used Across the Country to Predict Future Criminals. And it’s Biased Against Blacks.” *Federal probation*, *80*.

11. ANÁLISIS DE FLORES ET AL.

Como se mencionó antes, meses después de que se publicara el artículo original de ProPublica (Julia Angwin, 2016), se publicó un artículo con diversas objeciones a dicho estudio (W. Flores et al., 2016), argumentando que el sesgo alegado por ProPublica no estaba basado en un análisis adecuado.

Una de las objeciones argumenta que ProPublica interpreta, injustificadamente, los puntajes COMPAS como un resultado binario, pues usan sólo puntajes altos y bajos, considerando los puntajes medios como altos, y se alega que el estándar en este caso es estudiar también los cambios producidos al considerar los puntajes medios como bajos. Los autores no se dedicaron a poner esto a prueba, pero se hizo en el anexo de resultados adicionales de este escrito, donde se observó que sí hay cambios, pero no tienen ningún efecto sustancial en las conclusiones principales de ProPublica sobre sesgo racial. Otra crítica interesante es hacia el uso de la significación estadística en una parte del estudio original (diferente de las examinadas en este trabajo), denunciando que ProPublica no usó los estándares usuales para dicha significación, y sugiriendo que de cualquier modo alcanzar la significación estadística no es lo ideal para lograr un resultado significativo en dicho procedimiento. Esto es interesante porque de cualquier modo, al no usar la metodología bayesiana, los autores de esta crítica tuvieron que reportar la significación estadística en algunos de sus resultados (W. Flores et al., 2016, pp. 7–10).

El objetivo en este capítulo no es analizar las críticas de este segundo trabajo, o tratar de conciliar las diferencias entre las dos partes en este debate, sino observar los efectos de la metodología bayesiana en algunas conclusiones derivadas de esta crítica, como se hizo con el estudio de ProPublica. Para justificar la ausencia de sesgo racial en la herramienta COMPAS, los autores llevan a cabo dos procedimientos principales, y el que es de interés en este escrito es su análisis de la relación entre puntajes COMPAS y reincidencia, para el cual usan modelos de regresión logística.

Primero, como en el caso anterior, los modelos serán implementados en una reproducción del trabajo original, comparando los resultados originales con los obtenidos, y presentando las conclusiones de los autores. Posteriormente, se planteará el diseño de los modelos, y finalmente se pasa a la implementación usando la metodología bayesiana y al análisis de las diferencias en los resultados y conclusiones que se obtengan. Los detalles implícitos en la implementación, que ya se han expuesto en capítulos anteriores, y en particular en el análisis del estudio de ProPublica, no se revisarán de nuevo (aunque se mencionarán cuando sea pertinente), y los resultados se presentarán de una forma más directa, aunque también surgirán nuevas discusiones y detalles propios de este nuevo caso de estudio.

11.1 REPRODUCCIÓN

A diferencia de ProPublica, los autores de la réplica no hicieron disponible el código para sus implementaciones (W. Flores et al., 2016), y ni siquiera hacen referencia a las herramientas utilizadas, o a los detalles de los modelos (como el algoritmo de optimización o las funciones específicas usadas). Sin embargo, dicen haber usado la misma sintaxis que ProPublica (W. Flores et al., 2016, p. 10), así que posiblemente también usaron el lenguaje R y las mismas funciones.

Preprocesamiento de Datos

Los datos usados son exactamente los mismos que en el trabajo de ProPublica, pero hay una diferencia importante en la réplica: ya que el artículo de ProPublica sólo denuncia un prejuicio existente en COMPAS contra los elementos de raza negra, ante elementos de raza blanca, W. Flores et al. (2016) filtran los datos para hacer su estudio únicamente con elementos de estas dos razas. Conservando únicamente estos elementos, los datos para reincidencia general se reducen de 6172 a 5278 registros. El número de registros confirma que los datos fueron filtrados como en el análisis de

ProPublica, pues de haber considerado los elementos de estas dos razas antes de filtrar los datos con el criterio de dicho estudio, se tendrían 6150 elementos.

En esta réplica hay una breve etapa de exploración de datos, que es parte de la primera fase del trabajo de W. Flores et al. (2016), la cual no involucra modelos clasificadores y por lo tanto no se reproduce en esta tesis. Sin embargo, esta breve exploración se puede desarrollar fácilmente, usando la librería Pandas como en la reproducción de ProPublica. Los autores presentan las proporciones porcentuales que reincidieron, considerando raza y puntaje COMPAS. Los resultados obtenidos coinciden con los de los autores, usando la precisión numérica sin decimales que emplean para todos los datos (W. Flores et al., 2016, p. 12):

	Puntaje COMPAS categórico			
Población	*Cualquier puntaje*	*Bajo*	*Medio*	*Alto*
Total	47%	32%	55%	75%
Raza negra	52%	35%	56%	75%
Raza blanca	39%	29%	53%	73%

Tabla 11.1. Porcentaje de elementos en la población que reincidieron, agrupando por puntaje y raza.

Como en el caso de ProPublica, los autores no derivan conclusiones directamente de la etapa de exploración de datos, pero sí hacen notar que a pesar de que considerando todos los elementos los de raza afroamericana reinciden en una mayor proporción, las proporciones entre razas tienen una mayor similitud al comparar por puntajes categóricos. Los autores también puntualizan que las diferencias raciales en la reincidencia de los individuos dependen del comportamiento de los mismos y del sistema de justicia, no de la herramienta predictiva.

El Modelo Clasificador

En esta fase se explora la relación entre el puntaje COMPAS y la reincidencia, mediante varios modelos de regresión logística. En el estudio de ProPublica se usó un solo modelo, y su salida se empleó para calcular medidas de la discriminación. En este caso, los autores implementaron cuatro modelos, usando distintas variables como entradas, y argumentan la inexistencia del sesgo con base en las diferencias entre los distintos modelos, no con base en los valores de medidas de discriminación en sí. Los atributos usados en este caso ya no son exclusivamente categóricos, y se explican a continuación:

Age (edad): esta es una variable discreta, que expresa la edad de los individuos, en años, con un número entero.

Female (género femenino): es una variable binaria que indica si un individuo es de género femenino. Es cero cuando el individuo es de género masculino.

Black (raza negra): esta variable binaria indica si un individuo es de raza afroamericana, y es cero para elementos de raza caucásica.

NPC Decile (puntaje COMPAS): es el puntaje COMPAS en su forma no categórica, es una variable discreta, que es un número entero de 1 a 10.

NPC Decile X Black (puntaje COMPAS X raza negra): este es un término de interacción entre raza negra y puntaje COMPAS. Para cualquier elemento de raza afroamericana, simplemente es el mismo valor de su puntaje COMPAS, y para elementos de raza caucásica su valor es cero. Por su definición es una variable discreta, que es un número entero de 0 a 10.

two_year_recid (reincidencia en dos años): es una variable binaria que indica si el individuo reincidió en los dos años posteriores a la asignación de su puntaje COMPAS.

En este caso la reincidencia, o el atributo *two_year_recid*, no el puntaje COMPAS, es la respuesta o salida (*y*), así que la probabilidad de éxito del modelo de regresión logística es la probabilidad de que un elemento reincida en dos años. Esta salida no requiere modificaciones para usar la regresión logística, porque ya es binaria en los datos originales. El resto son las variables de entrada, o componentes del vector **x** en la tarea de clasificación. Al igual que la salida, no hace falta modificar atributos de edad ni puntaje COMPAS, pues en los datos originales ya tienen la forma necesaria. En cuanto a la raza y el género, como en el capítulo anterior, se les hace la codificación "one-hot" para que sean realmente variables binarias; como ambas variables tienen sólo dos categorías cada una, esta codificación las convierte en variables binarias simples (en contraste con la raza en el capítulo anterior, que era representada por 5 variables binarias porque tenía 6 categorías posibles). Finalmente el término de interacción entre puntaje COMPAS y raza negra se obtiene simplemente multiplicando la variable binaria para raza negra y el puntaje COMPAS para cada elemento.

Los autores usan cuatro modelos para estudiar la relación entre puntaje COMPAS y reincidencia, los cuatro modelos usan el atributo de reincidencia en dos años como salida, pero cambian las covariables de entrada (componentes del vector **x**) a usar en cada modelo, pues el análisis consiste en observar las diferencias al descartar o añadir algunas de estas. Todos los modelos usan los atributos de edad y género femenino, el atributo de raza negra se usa en todos los modelos menos el segundo, el de puntaje COMPAS se usa en todos menos el primero, y finalmente el término de interacción sólo se usa en el cuarto modelo, que es el único que usa todos los atributos de entrada. Para mayor claridad esta información se resume a continuación:

	Atributos de entrada (componentes del vector **x**)				
Modelo	*Edad*	*Género femenino*	*Raza negra*	*Puntaje COMPAS*	*Puntaje COMPAS X Raza negra*
1	X	X	X		
2	X	X		X	
3	X	X	X	X	
4	X	X	X	X	X

Tabla 11.2. Atributos de entrada empleados en cada uno de los cuatro modelos de regresión logística.

En este caso, los pesos **w** y la constante *b* para cada modelo, se obtienen usando la misma función que en la reproducción del estudio de ProPublica. Esta función es "tfp.glm.fit" en TensorFlow Probability, que en el capítulo pasado dio los mismos resultados que la función "glm.fit" en R (*Tfp.Glm.Fit | TensorFlow Probability*, s/f), (*Glm.Fit Function | R Documentation*, s/f). Con esta reproducción, se obtuvieron los coeficientes para los cuatro modelos.

Coeficientes de Regresión Logística (componentes del vector **w** y valor de *b*)				
	Modelo			
Atributo correspondiente en el vector **x**	*Modelo 1*	*Modelo 2*	*Modelo 3*	*Modelo 4*
Ninguno (el coeficiente es la constante *b*)	0.82784	-0.86296	-0.91222	-0.93188
Edad (A*ge*)	-0.03131	-0.01131	-0.01093	-0.01083
Género femenino (F*emale*)	-0.53685	-0.50385	-0.49795	-0.49926
Raza negra (*Black*)	0.37233	N/A	0.08312	0.11056
puntaje COMPAS (*NPC Decile*)	N/A	0.2638	0.26051	0.26473
Puntaje COMPAS X raza negra (*NPC Decile X black*)	N/A	N/A	N/A	-0.00632

Tabla 11.3. Coeficientes obtenidos en la reproducción.

Los coeficientes no se pueden comparar porque los autores no presentan sus resultados para estos. Los resultados que los autores presentan para cada atributo coinciden con el valor de una medida de la discriminación, el radio de posibilidades (OR), para cada uno de ellos. Sin embargo, los autores usan la interpretación general de esta medida, que no siempre se usa para analizar la discriminación. Esta forma general de interpretar el OR se discute más adelante, y es con base en ella que los autores expresan sus conclusiones para los datos presentados. La comparación entre reproducción y trabajo original debe hacerse directamente con el OR para los atributos de los cuatro modelos.

Se mencionó que en el trabajo de ProPublica se reporta el criterio AIC, como una validación de los modelos ajustados, pero no se considera realmente para las conclusiones ni el análisis (de hecho este criterio es calculado automáticamente por la función del lenguaje R que usan, así que reportarlo no fue necesariamente una decisión premeditada de ProPublica). En este caso, los autores reportan tres medidas similares para cada modelo: la prueba de chi cuadrada, la verosimilitud logarítmica (LL o *log-likelihood*) y la medida pseudo-R^2. Como ya se ha mencionado antes, la NLL es la función de costo a minimizar, y simplemente es el negativo de la LL, que como ya se ha mencionado, es una función de los parámetros **w**, indicando la probabilidad de que éstos expliquen los datos observados[83]. Ya que la NLL es la función que se minimiza durante el entrenamiento, la LL se maximiza, y es una medida de cuán bueno es el ajuste del modelo a los datos. Los valores para la prueba de chi cuadrada y la pseudo R^2, para algún modelo, se obtienen con expresiones que comparan la LL del modelo y la LL de un modelo "nulo", que trata de explicar los datos sin usar atributos aparte de la constante *b* (como el que se usa en el calculo del AIC)[84].

La diferencia con ProPublica, es que en este caso los autores sí usan estas medidas para derivar conclusiones sobre el sesgo racial, por lo que los resultados obtenidos para estas medidas se presentan en este capítulo. Sin embargo, hay otras comparaciones de rendimiento, enfocadas a la generalización de los modelos, que no se tratan ni en este estudio ni en el de ProPublica, y las mismas se discuten en la sección correspondiente del anexo para ambos casos.

83 En particular, la expresión sería: LL = log [p (**y** | **X**, **w**)]. Donde p es la distribución de Bernoulli que usa la regresión logística.

84 Considerando que LL_m y LL_0 son las LL del modelo y del modelo nulo respectivamente, la expresión para chi cuadrada es: $\chi^2 = -2(LL_0 - LL_m)$. A su vez, la expresión para pseudo R^2 es: pseudo-$R^2 = 1 - (LL_m / LL_0)$ (*lectur21*, s/f), (Allison, s/f).

De esta forma se obtuvieron prácticamente los mismos resultados que en (W. Flores et al., 2016, p. 28), para la precisión de dos decimales que usan en su trabajo.

	Reproducción				**Flores et al.**			
	Modelo				*Modelo*			
Atributo	*1*	*2*	*3*	*4*	*1*	*2*	*3*	*4*
Edad	0.97	0.99	0.99	0.99	0.97	0.98	0.99	0.99
Género femenino	0.58	0.60	0.61	0.61	0.58	0.60	0.61	0.61
Raza negra	1.45	N/A	1.09	1.12	1.45	N/A	1.09	1.12
Puntaje COMPAS	N/A	1.30	1.30	1.30	N/A	1.30	1.30	1.30
Puntaje COMPAS X raza negra	N/A	N/A	N/A	0.99	N/A	N/A	N/A	0.99
Constante	2.29	0.42	0.40	0.39	2.29	0.42	0.40	0.39

Tabla 11.4.1. Resultados obtenidos para el OR de los atributos.

	Reproducción				**Flores et al.**			
	Modelo				*Modelo*			
Medida	*1*	*2*	*3*	*4*	*1*	*2*	*3*	*4*
Chi cuadrada	297.68	804.42	806.13	806.20	297.68	804.42	806.13	806.19
LL	-3500.37	-3247.00	-3246.14	-3246.11	-3500.37	-3247.00	-3246.14	-3246.11
Pseudo-R^2	0.04	0.11	0.11	0.11	0.04	0.10	0.11	0.11

Tabla 11.4.2. Resultados obtenidos para medidas de rendimiento de los modelos.

Como se puede comprobar en las tablas, al obtener el OR y las tres medidas de rendimiento, se obtienen resultados idénticos a los de W. Flores et al. (2016) en casi todos los casos (las excepciones son en el OR para el atributo de edad en el segundo modelo, y en la chi cuadrada para el cuarto modelo, pero la diferencia es mínima). Antes de proceder con las conclusiones que los autores derivan de estos resultados, es importante recordar el significado del OR, y puntualizar cómo en este contexto, la interpretación es diferente a la empleada a lo largo de la tesis hasta ahora. La definición expresada en el capítulo de análisis de la discriminación dicta que:

$$OR = \frac{RR}{RC} = \frac{\frac{1-p_1}{1-p_2}}{\frac{p_1}{p_2}} = \frac{\frac{1-p_1}{p_1}}{\frac{1-p_2}{p_2}}$$

En ese contexto p_1 y p_2 son la probabilidad de recibir un beneficio para el grupo protegido y el no protegido, respectivamente. Con esto en mente, el OR no es más que las posibilidades (*odds*) de recibir un resultado negativo para el grupo protegido, entre las posibilidades de recibir un resultado negativo para el grupo no protegido[85].

Para el modelo de regresión logística, cuando se trata exclusivamente con variables categóricas y codificadas, como en el estudio de ProPublica en el capítulo anterior, las

85 Un resultado "negativo" aquí quiere decir que tiene una connotación negativa, pero no es necesariamente un resultado binario negativo (cero lógico). De hecho, en los modelos implementados en este trabajo, los resultados negativos son positivos en el sentido numérico, como recibir un puntaje alto o manifestar reincidencia delictiva.

probabilidades p_1 y p_2 para un individuo de características específicas (hay que recordar que el OR siempre se calcula con un atributo en particular, para cuantificar diferencias entre elementos que lo poseen y elementos que no) se pueden obtener directamente de la probabilidad de Bernoulli que es la salida del modelo, mediante los coeficientes que se hayan aprendido para dichas características. Sin embargo, en un sentido general, la expresión matemática del OR es la pendiente en la relación entre las posibilidades de recibir el resultado (evento exitoso en el modelo de Bernoulli de la regresión logística) y cualquier atributo cuyo coeficiente se tome en cuenta. Una explicación más simple es: la expresión matemática del OR, para algún atributo, da como resultado la proporción en que aumentan las posibilidades de recibir el resultado estudiado (un uno lógico), al incrementar en una unidad el valor de dicho atributo[86].

Entonces, cuando se trata de variables binarias, el OR tiene el significado dado en análisis de la discriminación, representando la diferencia de posibilidades entre dos grupos. En este caso, para los atributos de género y raza, la interpretación sigue siendo la misma del capítulo anterior: es la proporción en que aumentan las posibilidades de reincidencia para elementos de género femenino en el primer caso (en comparación con elementos similares de género masculino), y en el segundo es la proporción en que aumentan para elementos de raza afroamericana (en comparación con elementos similares de raza caucásica). Sin embargo, para variables no binarias, como la edad, el puntaje COMPAS, o la interacción entre éste y el atributo de raza negra, es mejor interpretarlo como la pendiente o aumento para cada unidad en que incrementa la variable. Es decir, para la edad, el OR representa la proporción en que aumentan las posibilidades con cada año de incremento, y para puntajes COMPAS, la proporción en que lo hacen con cada punto añadido.

Ahora bien, con esta interpretación aclarada, los autores usan los resultados para justificar la herramienta COMPAS, diciendo que en primer lugar, la comparación entre los modelos 3 y 4 muestra que añadir el término de interacción entre puntaje COMPAS y raza negra no tiene efectos significativos ni mejora la capacidad de predecir la reincidencia, y por lo tanto la relación entre el puntaje COMPAS y la reincidencia es la misma para elementos de razas afroamericana y caucásica. Esto se puede notar en el hecho de que el OR para el término de interacción es casi unitario (su valor es 0.99), por lo que aumentar una unidad en este atributo hace que las posibilidades se reduzcan sólo en un 1%. En cuanto a la mejora en la predicción de la reincidencia, los cambios para las medidas de chi cuadrada, LL y pseudo-R^2, son prácticamente nulos, así que aunque técnicamente los valores de chi-cuadrada y LL indican una mejora, es demasiado pequeña para considerar que es significativa. También se puede notar que los cambios para el resto de los resultados de los OR, entre estos dos modelos, son muy reducidos.

En segundo lugar, los autores argumentan que al comparar los modelos 2 y 3, no se observan diferencias raciales significativas en el término constante en la relación entre puntaje COMPAS y reincidencia. Es cierto que tanto en la tabla de coeficientes como en los OR, el cambio para la constante *b* es pequeño. El cambio en la constante se puede interpretar como la diferencia en las posibilidades del grupo general representado por la constante *b* al excluir del mismo a los miembros de raza negra. Estrictamente hablando, la ligera disminución en el OR para la constante (de 0.42 a 0.40), significa que cuando el grupo general no incluye miembros de raza negra, las posibilidades disminuyen, lo cual es coherente con el valor de 1.09 para el OR de raza negra en el modelo 3, que indica un ligero incremento en las posibilidades para miembros de raza negra.

Finalmente los autores concluyen que no hay diferencias significativas en la forma funcional de la relación entre puntajes COMPAS y reincidencia para elementos afroamericanos y caucásicos, por lo que un puntaje COMPAS se traduce en una probabilidad de reincidencia similar para ambas razas (W. Flores et al., 2016, pp. 15–16). Para ilustrar esta relación, los autores usan la probabilidad de reincidencia estimada por el modelo 4 para cada elemento, agrupando los resultados por puntaje

86 Una explicación sencilla e interactiva, que puede servir para aclarar más el concepto de la “pendiente” asociada con el OR, se halla en (*Simple Logistic Regression*, s/f).

COMPAS y por raza. Esta gráfica también se reproduce y coincide con la presentada en el trabajo original (W. Flores et al., 2016, Fig. 1.).

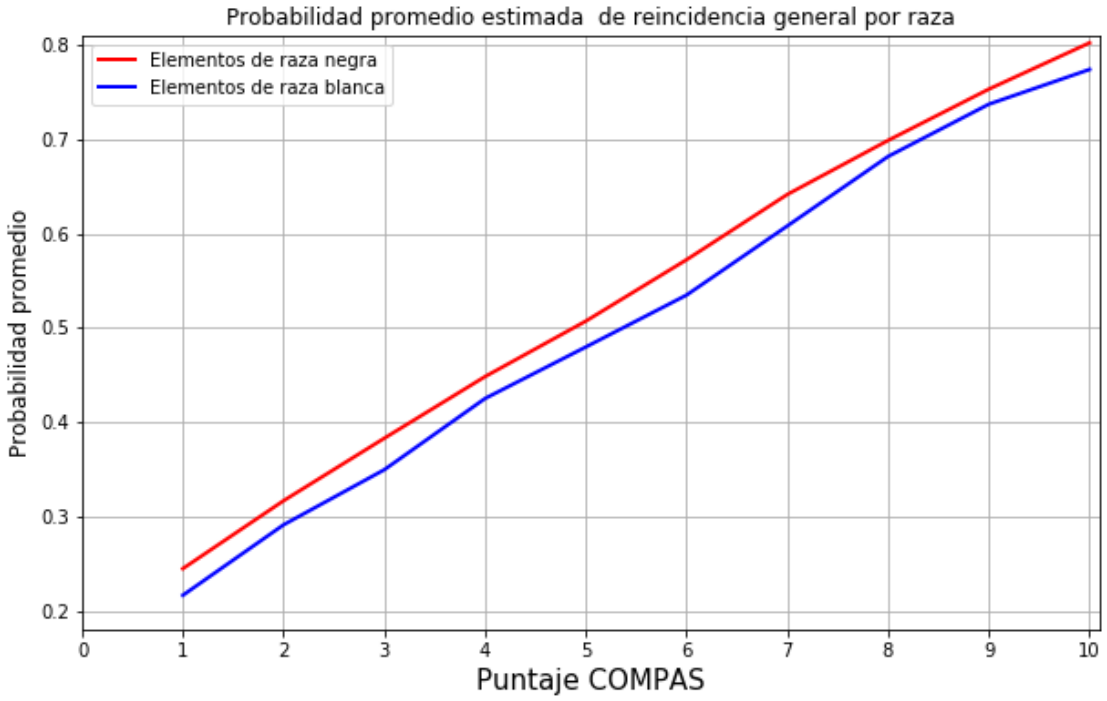


Figura. 11.1.1. Relación entre puntaje COMPAS y probabilidad estimada de reincidencia, para razas negra y blanca.

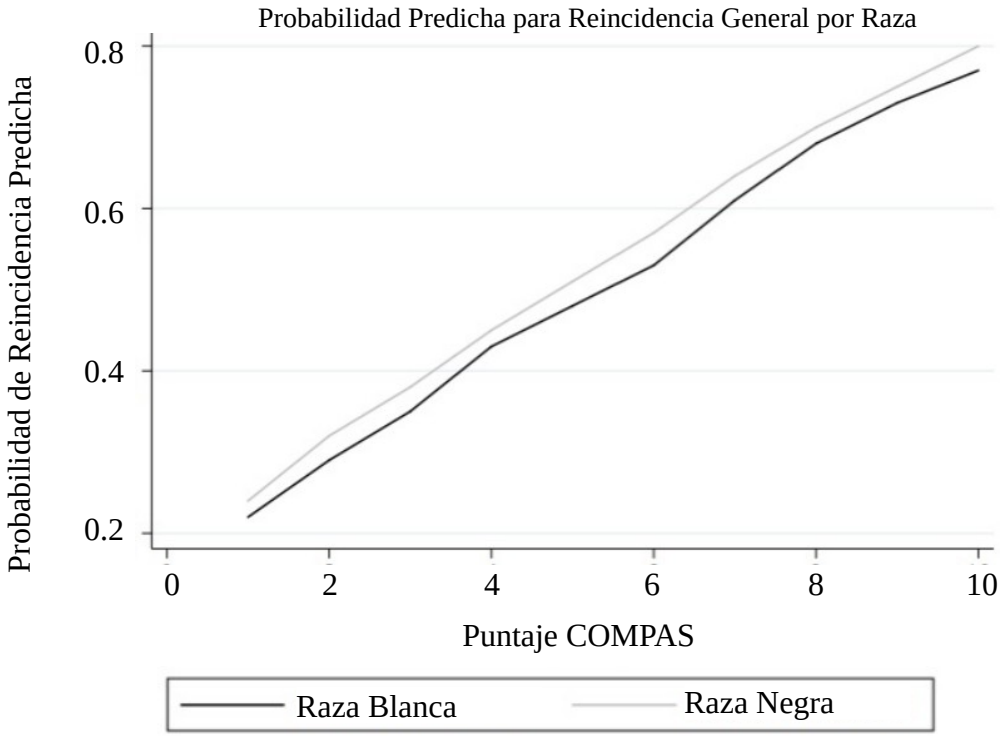


Figura. 11.1.2. Relación entre puntaje COMPAS y probabilidad estimada de reincidencia, tomada y traducida de (W. Flores et al., 2016, Fig. 1.).

Debido a la similitud entre las pendientes graficadas, los autores concluyen que a pesar de que se predigan y se observen más arrestos para la raza negra, la relación entre puntajes COMPAS y reincidencia no varía entre razas.

Reincidencia Violenta

Como en el capítulo anterior, ahora se presentan los resultados para reincidencia violenta, sin ahondar en los detalles de cada punto del proceso, pues ya se desarrollaron en la sección de reincidencia general. Una discrepancia importante en el trabajo de W. Flores et al. (2016) es el preprocesamiento de datos para reincidencia violenta. Como se vio en el caso general, los 5278 datos que usan los autores, en efecto son los datos de razas negra y blanca tras llevar a cabo el mismo preprocesamiento del trabajo de ProPublica; sin embargo, para reincidencia violenta los autores reportan que después de excluir elementos de razas diferentes a la caucásica y la afroamericana, quedan 3967 registros. Esto es problemático, porque al reproducir esta parte del proceso se halló que los registros para estas dos razas son efectivamente 3967, pero sólo si se toman antes de realizar el filtrado implementado por ProPublica, ya que si se filtraran los datos como en dicho estudio, y después se separaran los elementos de estas dos razas, el número de registros es 3377. Los autores nunca hacen referencia a esta discrepancia ni dan razones para filtrar casos de reincidencia general y no de reincidencia violenta, de hecho, considerando que afirman usar la misma sintaxis del estudio de ProPublica, lo más posible es que esto haya sido un error (W. Flores et al., 2016, p. 10).

Como se verá en los resultados, esta discrepancia se propaga a los procesos restantes, pues todos ellos dependen de los datos empleados. En esta reproducción se implementan los procedimientos usando los datos que supuestamente usaron los autores (los 3967 registros que se obtienen sin filtrar los datos), y también con los datos que deberían usarse considerando el procedimiento de reincidencia general (los 3377 registros, filtrados como en el trabajo de ProPublica).

La exploración se desarrolla de la misma manera que para la reincidencia general, pero ya que en este caso los resultados de la reproducción no coinciden con los de los autores, se presenta también la comparación con los mismos. Los cuatro modelos implementados son iguales a los que se usaron para reincidencia general, pero usan los datos y puntajes de reincidencia violenta, y los coeficientes se obtienen con los mismos métodos. Los resultados son los siguientes:

	Reproducción								**Flores et al.**			
	Datos Filtrados (N = 3377)				*Datos sin Filtrar (N = 3967)*							
Población	*Cualquier puntaje*	*Bajo*	*Medio*	*Alto*	*Cualquier puntaje*	*Bajo*	*Medio*	*Alto*	*Cualquier puntaje*	*Bajo*	*Medio*	*Alto*
Total	17%	9%	21%	45%	17%	9%	21%	41%	17%	11%	26%	45%
Raza negra	21%	10%	24%	45%	21%	10%	23%	42%	21%	13%	27%	47%
Raza blanca	12%	8%	16%	42%	12%	8%	18%	36%	12%	9%	22%	38%

Tabla 11.5. Porcentaje de elementos en la población que reincidieron, agrupando por puntaje y raza (reincidencia violenta).

Coeficientes de Regresión Logística (componentes del vector **w** y valor de *b*)								
	Modelo							
Atributo correspondiente en el vector **x**	*Modelo 1*		*Modelo 2*		*Modelo 3*		*Modelo 4*	
	Datos Filtrados (N = 3377)	*Datos sin Filtrar (N = 3967)*	*Datos Filtrados (N = 3377)*	*Datos sin Filtrar (N = 3967)*	*Datos Filtrados (N = 3377)*	*Datos sin Filtrar (N = 3967)*	*Datos Filtrados (N = 3377)*	*Datos sin Filtrar (N = 3967)*
Ninguno (el coeficiente es la constante *b*)	-0.48211	-0.40862	-2.37853	-2.13964	-2.46571	-2.19796	-2.34961	-2.05650
Edad (A*ge*)	-0.03749	-0.03863	-0.01402	-0.01847	-0.01360	-0.01822	-0.01422	-0.01897
Género femenino (F*emale*)	-0.82114	-0.74542	-0.75635	-0.68789	-0.74665	-0.68037	-0.74031	-0.67233
Raza negra (*Black*)	0.50622	0.45013	N/A	N/A	0.16892	0.11798	0.01183	-0.07521
puntaje COMPAS (*NPC Decile*)	N/A	N/A	0.29295	0.26836	0.28587	0.26322	0.26448	0.23721
Puntaje COMPAS X raza negra (*NPC Decile X black*)	N/A	N/A	N/A	N/A	N/A	N/A	0.031394	0.03821

Tabla 11.6. Coeficientes obtenidos en la reproducción (reincidencia violenta).

	Reproducción								**Flores et al.**			
	Datos Filtrados (N = 3377)				*Datos sin Filtrar (N = 3967)*							
	Modelo				*Modelo*				*Modelo*			
Atributo	*1*	*2*	*3*	*4*	*1*	*2*	*3*	*4*	*1*	*2*	*3*	*4*
Edad	0.96	0.99	0.99	0.99	0.96	0.98	0.98	0.98	0.96	0.99	1.00	1.00
Género femenino	0.44	0.47	0.47	0.48	0.47	0.50	0.51	0.51	0.47	0.57	0.57	0.57
Raza negra	1.66	N/A	1.18	1.01	1.57	N/A	1.13	0.93	1.57	N/A	1.24	1.21
puntaje COMPAS	N/A	1.34	1.33	1.30	N/A	1.31	1.30	1.27	N/A	1.32	1.30	1.30
Puntaje COMPAS X raza negra	N/A	N/A	N/A	1.03	N/A	N/A	N/A	1.04	N/A	N/A	N/A	1.01
Constante	0.62	0.09	0.08	0.10	0.66	0.12	0.11	0.13	0.66	0.09	0.08	0.0

Tabla 11.7.1. Resultados obtenidos para el OR de los atributos (reincidencia violenta).

	Reproducción								**Flores et al.**			
	Datos Filtrados (N = 3377)				*Datos sin Filtrar (N = 3967)*							
	Modelo				*Modelo*				*Modelo*			
Medida	*1*	*2*	*3*	*4*	*1*	*2*	*3*	*4*	*1*	*2*	*3*	*4*
Chi cuadrada	162.79	396.02	398.51	399.15	183.53	423.89	425.34	426.46	183.53	345.34	350.49	350.52
LL	-1464.32	-1347.70	-1346.46	-1346.14	-1725.60	-1605.42	-1604.70	-1604.14	-1725.60	-1644.70	-1642.12	-1642.11
Pseudo-R^2	0.05	0.13	0.13	0.13	0.05	0.12	0.12	0.12	0.05	0.09	0.10	0.10

Tabla 11.7.2. Resultados obtenidos para medidas de rendimiento de los modelos (reincidencia violenta).

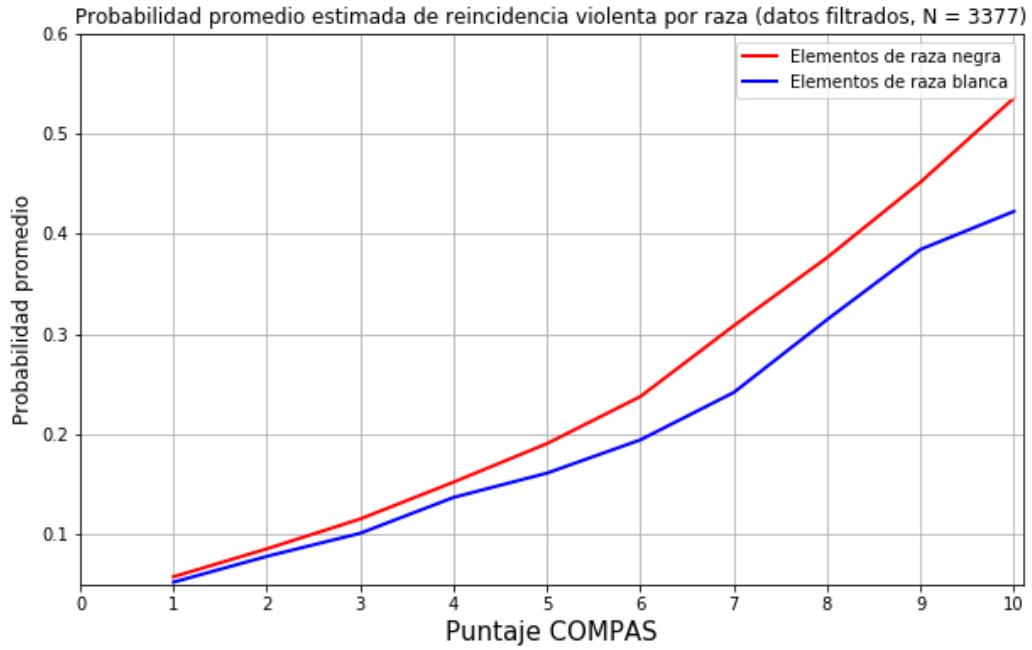


Figura 11.2.1. Relación entre puntaje COMPAS y probabilidad estimada de reincidencia violenta, para razas negra y blanca, usando los datos filtrados (3377 registros).

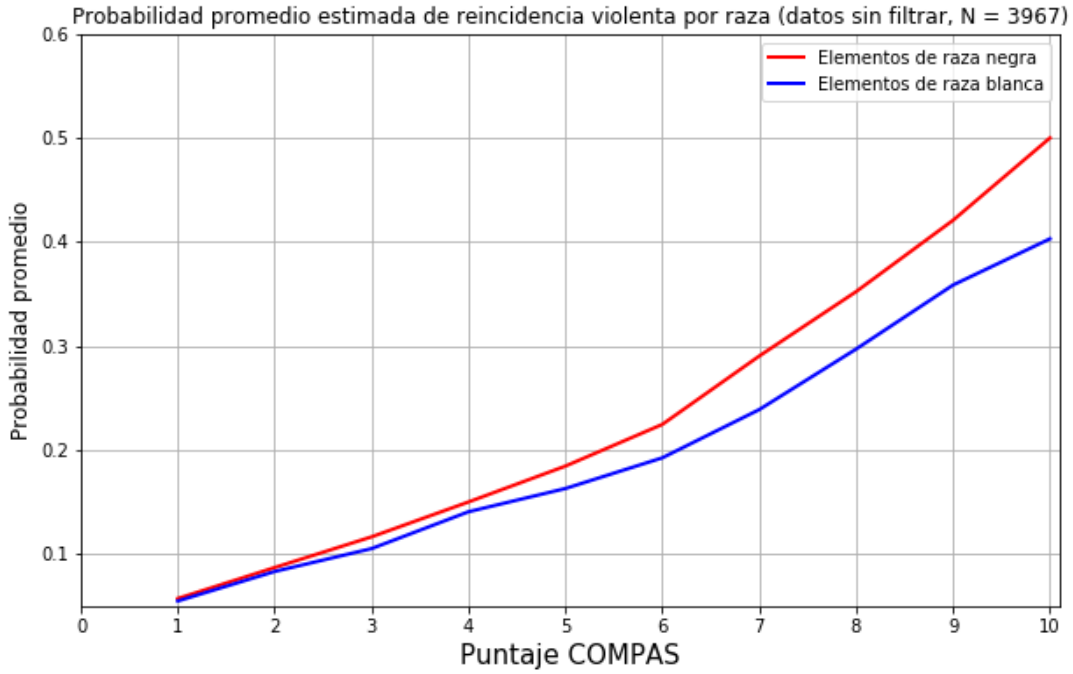


Figura 11.2.2. Relación entre puntaje COMPAS y probabilidad estimada de reincidencia violenta, para razas negra y blanca, usando los datos sin filtrar (3967 registros).

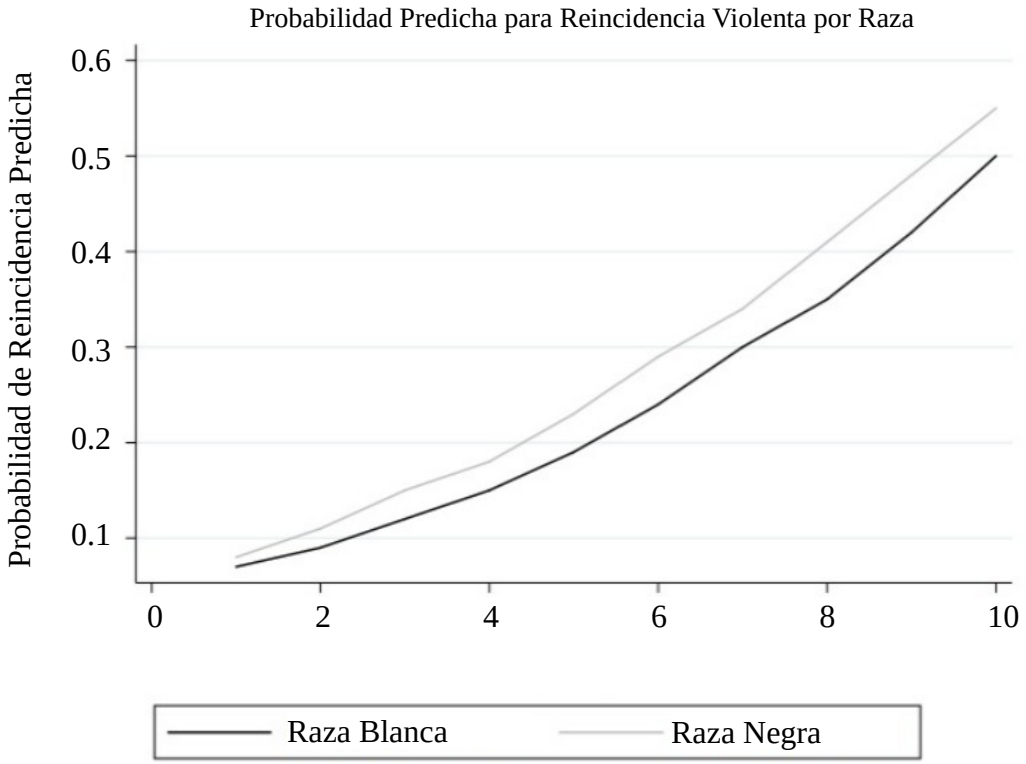


Figura 11.2.3. Relación entre puntaje COMPAS y probabilidad estimada de reincidencia violenta, tomada y traducida de (W. Flores et al., 2016, Fig. 2.).

Desde la exploración de los datos, se puede observar que el no filtrar los datos, como se hizo para la reincidencia general, afectará los resultados, ya que no sólo hay diferencias entre los resultados de la reproducción, sino que ninguno de los dos resultados, ya sea con o sin filtrado, coincide con los resultados presentados por los autores, poniendo en duda que hayan usado los datos originales de ProPublica para reincidencia violenta. Del mismo modo, para los coeficientes obtenidos, aunque las diferencias no son extremas al filtrar los datos, tampoco son cambios despreciables entre los modelos implementados. Como se mencionó antes, los autores no ponen disponibles los coeficientes obtenidos en su implementación, por lo que no se pueden comparar con los obtenidos en la reproducción como se hizo con la exploración de datos.

En las tablas para OR y medidas de rendimiento, nuevamente son evidentes las inconsistencias en el trabajo original. Para el modelo 1, los resultados de la reproducción, usando los datos sin filtrar, son idénticos a los presentados por los autores en ambas tablas, lo cual sugiere que entonces, como se esperaba, usaron los datos sin filtrar pero llevaron a cabo el mismo procedimiento. Sin embargo, para el resto de los modelos, los resultados ya no coinciden, e incluso los resultados de la reproducción para datos filtrados, se acercan más a los resultados de los autores en varios casos. Esto pone en evidencia que el trabajo de W. Flores et al. (2016) tiene algunas inconsistencias en su estudio de la reincidencia violenta, pues en ningún momento mencionan haber modificado el procedimiento usado para reincidencia general, y en general su análisis da a entender que el procedimiento es el mismo. Como los autores no publicaron su código, no es posible identificar cómo llegaron a estos resultados. A pesar de estas discrepancias, no suele haber variaciones extremas en los resultados de la reproducción. Los cambios más significativos para los OR son en los atributos de raza y género, en los modelos 2, 3 y 4. Y para las otras medidas, en general hay cambios evidentes en todos los modelos menos el primero.

Finalmente, las gráficas obtenidas en la reproducción no son muy diferentes entre sí, pero a diferencia de la gráfica presentada por los autores, estas gráficas muestran una divergencia entre las

funciones de cada raza, que va incrementando conforme aumenta el puntaje, mostrando que para puntajes más altos, la probabilidad incrementa en mayor medida para elementos de raza negra. Esta divergencia no se percibe en la gráfica presentada por los autores, y ahí la separación entre razas para puntajes COMPAS más bajos (menores a 5), es un poco más amplia que la observada en esta reproducción (W. Flores et al., 2016, Fig. 2.).

Con base en los resultados, los autores concluyen que, del mismo modo que en la reincidencia general, no se encuentran evidencias de sesgo racial, y en particular alegan que la relación entre raza y reincidencia violenta se vuelve insignificante una vez que el puntaje COMPAS forma parte del modelo. Es cierto que, como en el caso de reincidencia general, el cambio en las medidas de rendimiento es mínimo entre los modelos 3 y 4, aunque el cambio sigue favoreciendo al modelo 4 (esto es cierto aunque se filtren o no los datos). El OR para el término de interacción entre puntaje COMPAS y raza negra del modelo 4 aún tiene un valor cercano a la unidad, que indica un aumento de 1% en las posibilidades, por cada aumento unitario en el atributo, en los resultados de los autores, y aumentos un poco mayores de 3% y 4% para los resultados de la reproducción. En cuanto al cambio en el OR de la constante entre los modelos 2 y 3, también es cierto que el cambio es mínimo: tanto en los resultados de la reproducción como en el del trabajo original, este cambio es de tan sólo 0.01. Sin embargo, afirmar que la raza tiene un rol insignificante cuando se considera el puntaje COMPAS es algo precipitado, especialmente considerando los resultados de los autores, ya que en los modelos 3 y 4, donde están como atributos tanto la raza como el puntaje COMPAS, el atributo de raza negra tiene OR de 1.24 en el modelo 3 y de 1.21 en el modelo 4, y un aumento de más de 20% en las posibilidades no es precisamente insignificante. En los resultados de la reproducción tampoco se podría decir que la raza es insignificante si se considera el puntaje COMPAS, pues en el modelo 3 el OR sigue mostrando un aumento de posibilidades por encima del 10%, tanto para datos filtrados como para datos sin filtrar.

Finalmente, en la figura para probabilidades promedio por puntaje COMPAS en ambas razas, los autores afirman que al igual que en la reincidencia general, las pendientes son similares para reincidencia violenta, indicando que el aumento de probabilidad predicha de reincidencia violenta con el aumento de puntaje, no es diferente entre razas. Se puede decir que acorde a su gráfica (la figura 11.2.3 en este trabajo), es una conclusión aceptable, pero las gráficas obtenidas en esta reproducción, filtrando o no los datos, muestran que la probabilidad predicha tiene incrementos mayores, con cada aumento de puntaje, para la raza negra, especialmente para puntajes mayores a 4.

A diferencia de la reproducción para el trabajo de ProPublica, en este caso se encontraron inconsistencias en el trabajo de W. Flores et al. (2016), que fue necesario abordar en esta sección. Al presentar la implementación bayesiana, el enfoque estará en los cambios entre la reproducción y dicha implementación. Aunque no se ignorarán los resultados originales, el objetivo es dar énfasis a las implicaciones bayesianas, por lo tanto, en dicho capítulo no se hará una discusión tan profunda sobre las inconsistencias del trabajo original como en esta sección.

11.2 DISEÑO DE MODELOS

Como en el capítulo anterior, en esta sección se plantearán los modelos bayesianos, que mantienen las características descritas en dicho capítulo, pero los datos y la tarea de aplicación son diferentes. Los modelos no bayesianos usados en la reproducción de la sección anterior también se resumen con esquemas, al igual que el preprocesamiento de datos, pues a pesar de que se siguieron los pasos del capítulo anterior, es importante resaltar los cambios resultantes de las discrepancias para el estudio de la reincidencia violenta, sin mencionar que los atributos usados y la misma salida son diferentes en este caso.

Preprocesamiento

Como se expuso anteriormente, este punto no varía entre metodologías bayesiana y frecuentista. En este caso hay varios cambios en el preprocesamiento, incluyendo la inconsistencia mencionada para reincidencia violenta. La interpretación de los datos también es diferente, así que la entrada para los modelos de clasificación es distinta.

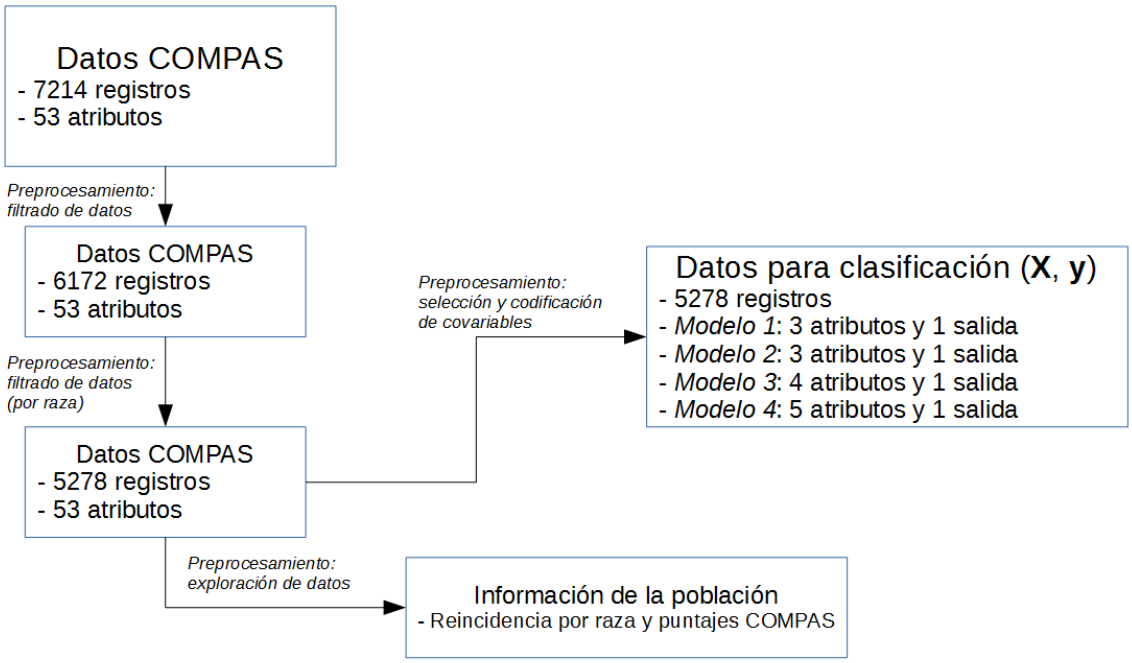


Figura 11.3.1. Esquema para el preprocesamiento de datos.

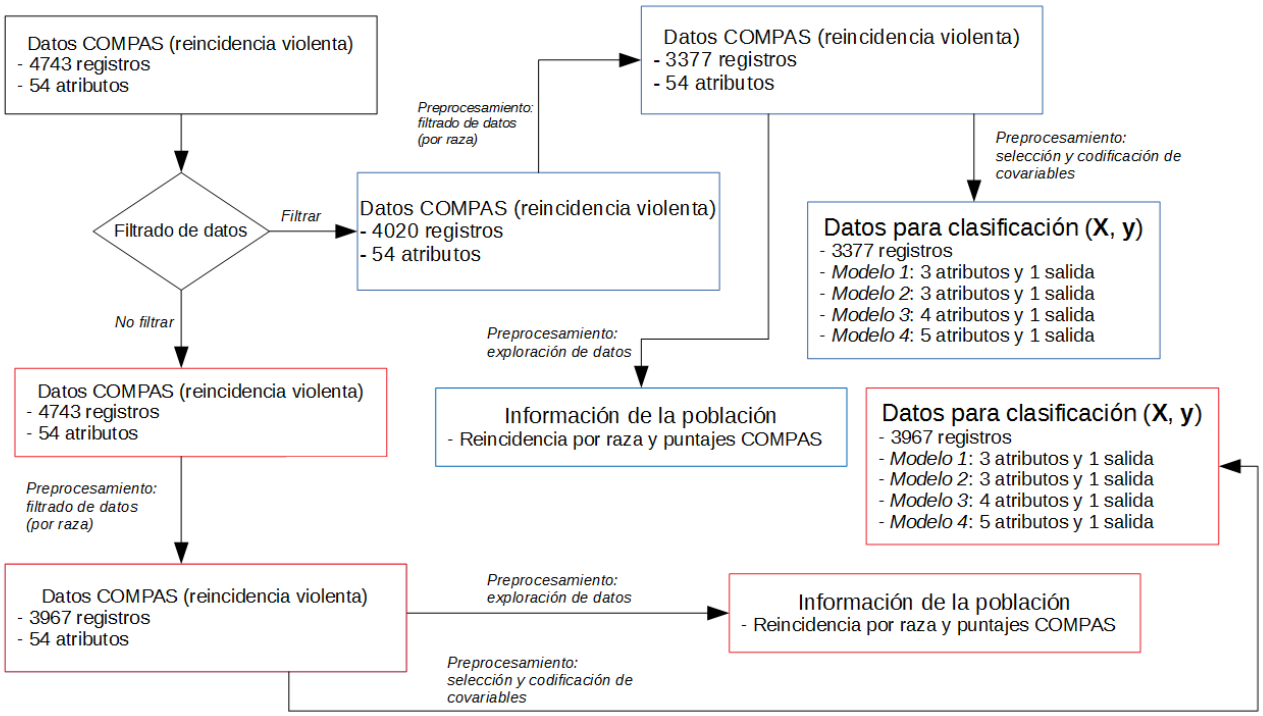


Figura 11.3.2. Esquema para el preprocesamiento de datos (reincidencia violenta).

El caso de reincidencia general es muy similar al presentado en el capítulo anterior, sólo se añade la parte de filtrado por razas, donde se quitan todos los registros que no sean de razas negra o blanca, y

la información obtenida en la exploración de datos tiene una forma distinta que en el caso de ProPublica. Del mismo modo, los datos de clasificación ahora contemplan cuatro interpretaciones diferentes, pues aunque los cuatro modelos usan los mismos datos, los atributos usados en cada caso son diferentes.

Para reincidencia violenta, las discrepancias en el filtrado de datos se representan en el esquema: el flujo es esencialmente el mismo que para reincidencia general, pero debe considerarse el caso donde se realiza el filtrado inicial de los datos y el caso contrario. Esto provoca resultados distintos para la información de la población y también en dos conjuntos de datos para clasificación, cada uno con sus cuatro interpretaciones diferentes dependiendo del modelo.

Modelado

Los modelos siguen teniendo las características descritas en el capítulo anterior, por lo que la representación es la misma, pero usando datos diferentes en la tarea de clasificación. También se debe mencionar que a pesar de tener cuatro modelos diferentes para los mismos datos, las características de los modelos son las mismas, pues finalmente no dependen de la tarea de clasificación.

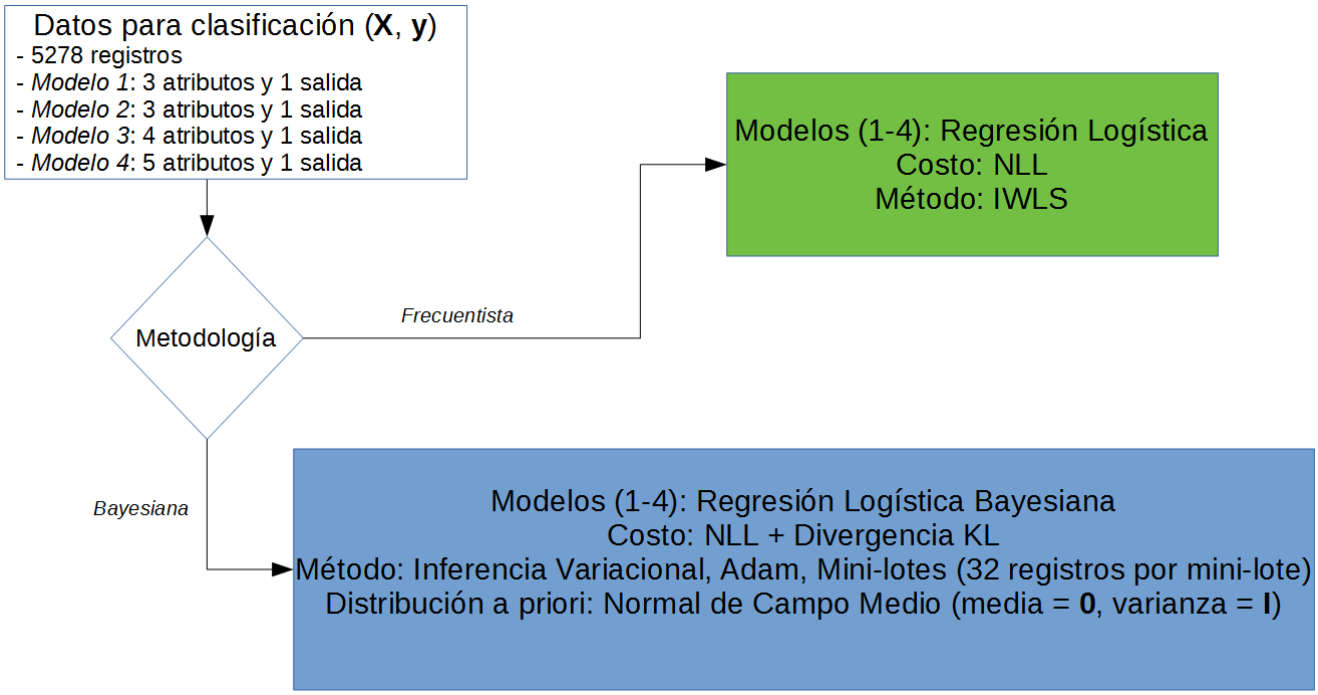


Figura 11.4.1. Esquema para la selección e implementación de metodologías y modelos.

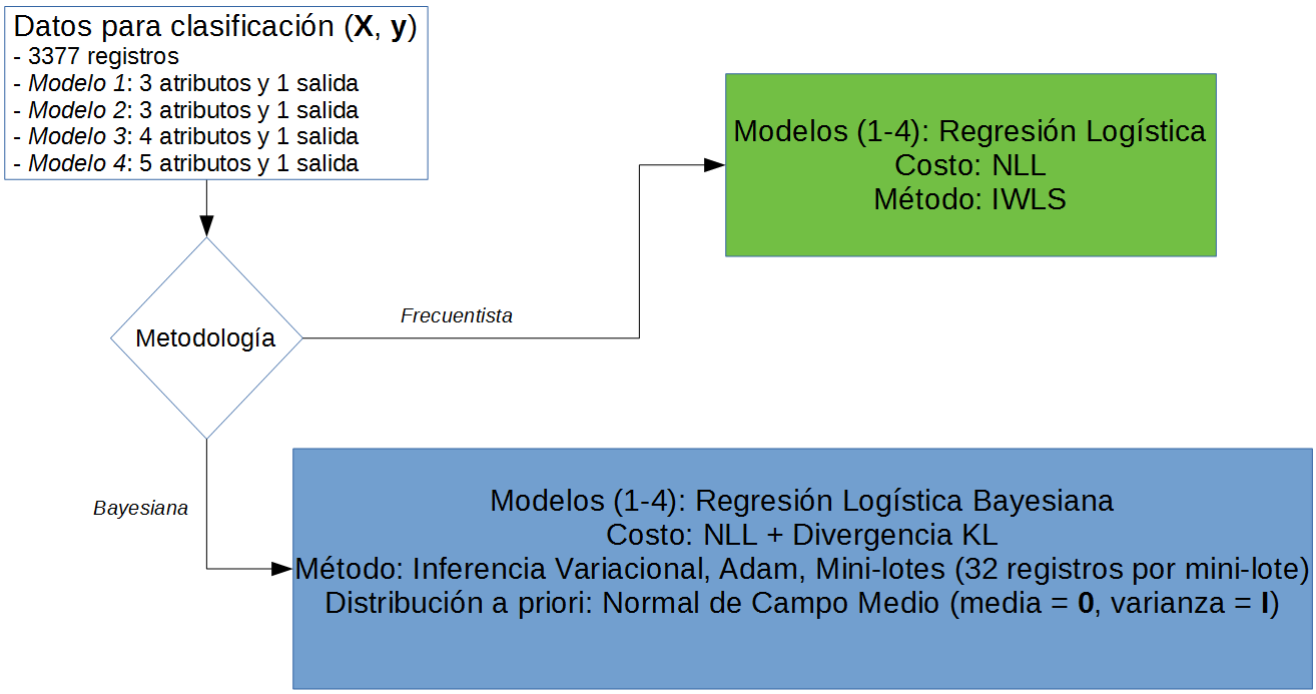


Figura 11.4.2. Esquema para la selección e implementación de metodologías y modelos (reincidencia violenta, filtrando datos iniciales).

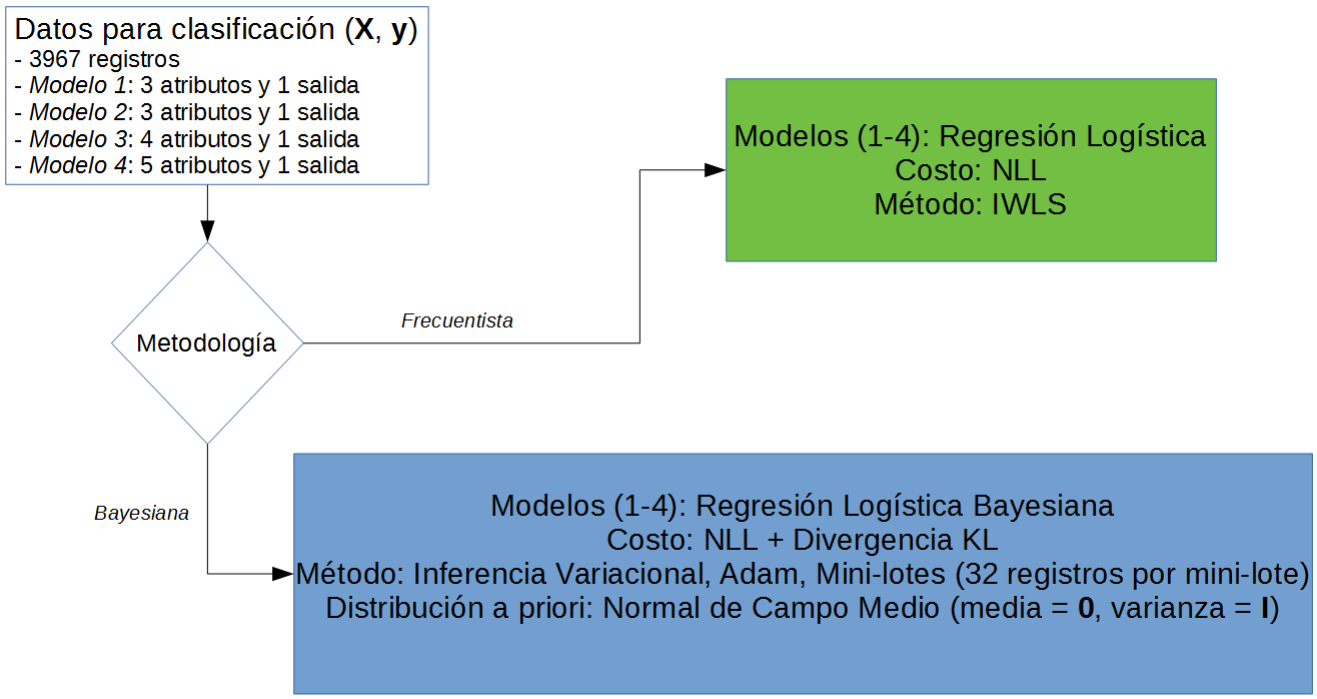


Figura. 11.4.3. Esquema para la selección e implementación de metodologías y modelos (reincidencia violenta, filtrando datos iniciales).

En los esquemas se puede comprobar cómo la única diferencia para el modelado son los datos de clasificación usados de los tres conjuntos de datos posibles (reincidencia general, reincidencia violenta y reincidencia violenta sin el filtrado de datos inicial). En los tres casos también se ilustra que las características de los modelos no cambian entre las cuatro interpretaciones (correspondientes a cada uno de los modelos del 1 al 4): nuevamente la única diferencia es entre las interpretaciones mismas, que finalmente siguen cambiando solamente los datos usados para la

clasificación. La diferencia es que entre los cuatro modelos, los atributos usados son los que cambian, no los registros (como sí se da al cambiar entre los tres conjuntos de datos mencionados).

No tiene caso cubrir nuevamente los detalles de estos modelos, que se han revisado en capítulos anteriores y en especial en la versión de esta misma sección para el análisis de ProPublica, así que sólo queda mostrar el esquema para la obtención de resultados.

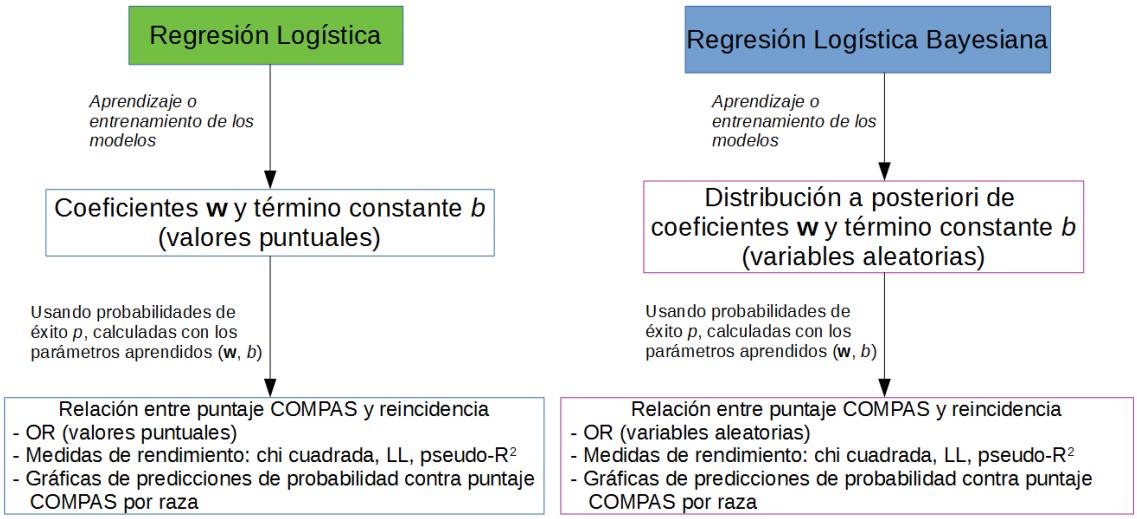


Figura. 11.4.4. Esquema para la obtención de resultados.

Como en el capítulo anterior, esta parte del proceso no depende de los datos usados en un sentido conceptual, así que el esquema es común para el caso de reincidencia general y los dos posibles casos para reincidencia violenta. Del mismo modo, nótese cómo los cuatro modelos para cada caso se integran en un mismo bloque, ya que los resultados obtenidos en cada caso integran el uso de los cuatro. Sin embargo, cabe resaltar que el modelo 4 es el único empleado para la obtención de las gráficas de predicciones de probabilidad contra puntaje COMPAS, como se mencionó en la reproducción desarrollada en la sección anterior.

Nuevamente, la diferencia entre métodos llevará a resultados en términos de variables aleatorias con distribuciones de probabilidad para el caso bayesiano, y su efecto podrá apreciarse en la siguiente sección. Sin embargo, nótese que en las medidas de rendimiento no se distingue entre valores puntuales o variables aleatorias entre métodos: al hacer predicciones con el método bayesiano, la media de la distribución a posteriori tiene los parámetros óptimos (*MLPR w6c - Machine Learning and Pattern Recognition*, s/f), por lo que se usan los valores de dicha media, resultando en medidas puntuales como en el caso frecuentista.

11.3 IMPLEMENTACIÓN BAYESIANA

En esta sección se presentan los resultados al sustituir los modelos de regresión logística por sus contrapartes bayesianas, implementadas con inferencia variacional. La mayoría de los detalles con respecto a los modelos descritos en el capítulo anterior son los mismos en este caso, y se debe asumir que se usaron los mismos procedimientos, a menos que se diga explícitamente lo contrario.

Coeficientes de Regresión Logística (componentes del vector **w** y valor de *b*)										
	Distribuciones a priori		*Distribuciones a posteriori*							
	Modelo		Modelo							
Atributo correspondiente en el vector **x**	*Todos los Modelos*[87] *(1-4)*		*Modelo 1*		*Modelo 2*		*Modelo 3*		*Modelo 4*	
	Media	Varianza	Media	Varianza	Media	Varianza	Media	Varianza	Media	Varianza
Ninguno (el coeficiente es la constante *b*)	0.0	1.0	0.76141	0.000823	-0.720139	0.00083	-0.719695	0.000608	-0.643705	0.000931
Edad (A*ge*)	0.0	1.0	-0.040205	2.4e-05	-0.020781	2.5e-05	-0.020596	2.1e-05	-0.023717	2.4e-05
Género femenino (F*emale*)	0.0	1.0	-0.54211	0.006393	-0.538809	0.005991	-0.530179	0.01056	-0.521103	0.004627
Raza negra (*Black*)	0.0	1.0	0.409263	0.000743	N/A	N/A	0.071843	0.00144	-0.054682	0.001842
puntaje COMPAS (*NPC Decile*)	0.0	1.0	N/A	N/A	0.254528	7.6e-05	0.25305	7.7e-05	0.231204	6.6e-05
Puntaje COMPAS X raza negra (*NPC Decile X black*)	0.0	1.0	N/A	N/A	N/A	N/A	N/A	N/A	0.041535	0.000125

Tabla 11.8. Distribuciones de los coeficientes en la implementación bayesiana.

Como en el capítulo anterior, se puede notar que hay diferencias entre las medias a posteriori de los coeficientes, y los valores hallados con el método frecuentista, aunque estas diferencias no son extremas. Como se mencionó para el caso de ProPublica, los cambios entre esquemas de optimización son una causa importante de las diferencias, y podrían tenerse resultados más parecidos si la misma optimización del método bayesiano se empleara con la metodología frecuentista.

A continuación se ilustran las distribuciones a posteriori para los coeficientes de los cuatro modelos por separado, y también la distribución a priori común para todos los coeficientes en todos los modelos.

87 Ya que en los cuatro modelos todos los coeficientes usan media de cero y varianza unitaria, esta información se resume en esta columna, sin embargo no debe perderse de vista que algunos coeficientes no están presentes en algunos modelos, así que está implícito que estas distribuciones a priori de los coeficientes sólo aplican en los modelos donde aparece cada uno.

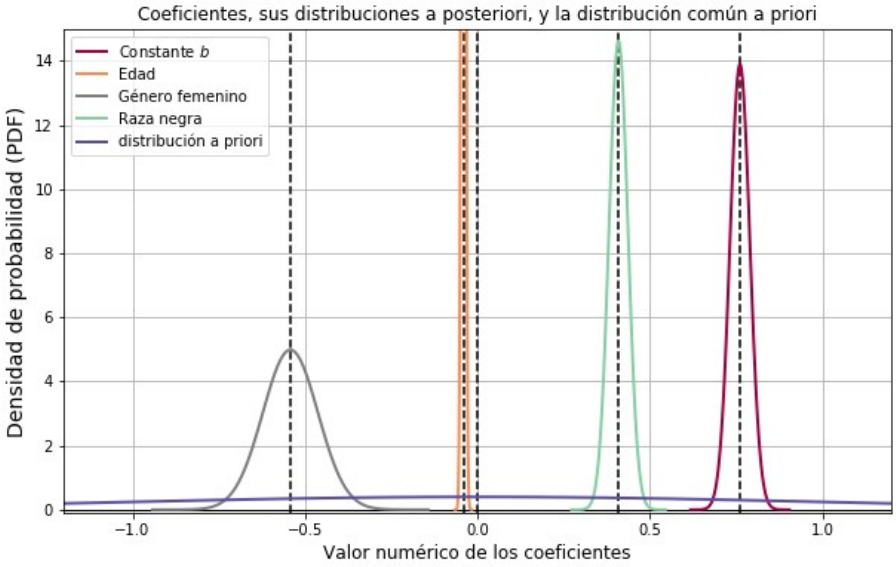


Figura. 11.5.1. Distribuciones para los coeficientes de la regresión logística bayesiana (modelo 1).

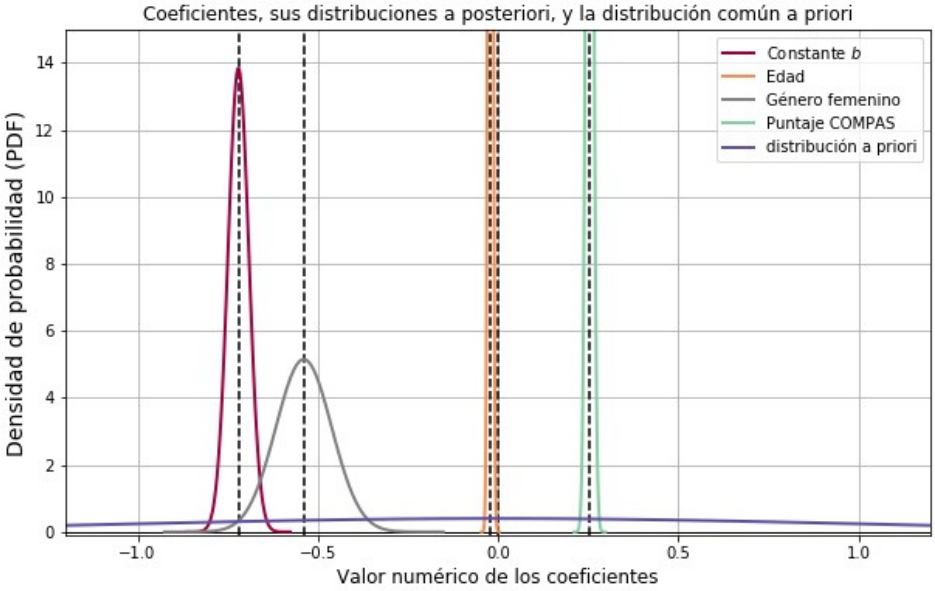


Figura. 11.5.2. Distribuciones para los coeficientes de la regresión logística bayesiana (modelo 2).

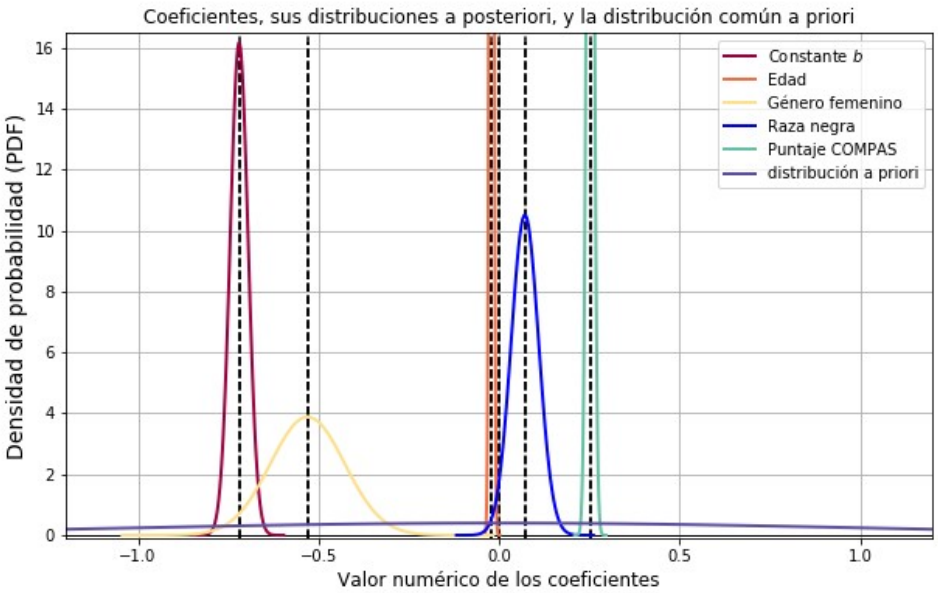


Figura. 11.5.3. Distribuciones para los coeficientes de la regresión logística bayesiana (modelo 3).

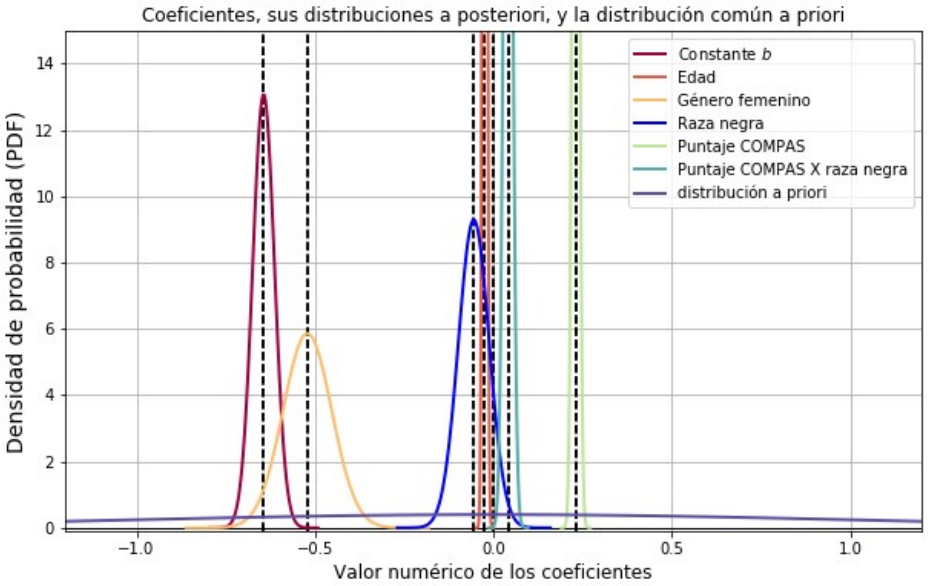


Figura. 11.5.4. Distribuciones para los coeficientes de la regresión logística bayesiana (modelo 4).

En este caso, ningún coeficiente, en ningún modelo, tiene varianzas elevadas en comparación a los coeficientes en el caso de ProPublica, esto ya podía esperarse de los resultados para las varianzas en la tabla 11.8. Hay muy poca incertidumbre para los coeficientes de edad, puntaje COMPAS, y la interacción entre puntaje COMPAS y raza negra, y también suelen tener medias muy similares de modelo a modelo.

Un punto interesante en estos resultados es la relación entre la incertidumbre y la interpretación de los datos. El coeficiente de género femenino es el que tiene mayor incertidumbre en todos los modelos, pero su media varía poco entre modelos. En cambio, la constante *b* y el coeficiente de raza negra tienen variaciones más notables en la media, aunque su varianza sea siempre menor a la del coeficiente para género femenino. Esto revela algunas características de la

relación entre la incertidumbre y la interpretación de los datos (la tarea de clasificación específica a resolver): aunque haya muchos datos para aprender alguna cualidad, si se cambia la interpretación de los datos, la respuesta referente a dicha cualidad también puede cambiar, manteniendo, antes y después del cambio, una certeza elevada para la respuesta. Por otro lado, el coeficiente de género femenino refleja lo contrario, pues aunque una característica tenga un efecto similar bajo diversas interpretaciones de los datos, esto no significa que se tenga certeza con respecto al valor que representa dicho efecto, por ello la media difiere poco entre los cuatro modelos, pero aún así la varianza es más elevada que la de otros coeficientes, cuyas medias varían más entre modelos.

Como una observación final, en este caso la varianza unitaria a priori claramente es más que suficiente para los cuatro modelos, pues en todos los casos su densidad de probabilidad se extiende bastante más allá de los valores hallados para los coeficientes.

A continuación se presentan los resultados para el radio de posibilidades (OR) y las medidas de rendimiento. El OR fue obtenido mediante el método de muestreo que se describió para las medidas de discriminación en el capítulo anterior, y también se obtuvieron los límites superior e inferior para este caso, usando los percentiles descritos en ese mismo capítulo. Como ya se comentó, las medidas de chi cuadrada, LL y pseudo-R^2, se obtienen con base en la media de las distribuciones a posteriori, sin considerar varianzas ni rangos.

	Modelo							
Atributo	*1*		*2*		*3*		*4*	
	Media	Varianza	Media	Varianza	Media	Varianza	Media	Varianza
Edad	0.960563	2.2e-05	0.979404	2.4e-05	0.979632	2e-05	0.976564	2.3e-05
Género femenino	0.58322	0.002162	0.585036	0.002038	0.591424	0.003732	0.595348	0.001646
Raza negra	1.506103	0.001678	N/A	N/A	1.075284	0.001659	0.948115	0.001659
puntaje COMPAS	N/A	N/A	1.289857	0.000125	1.287973	0.000127	1.260164	0.000104
Puntaje COMPAS X raza negra	N/A	N/A	N/A	N/A	N/A	N/A	1.042517	0.000136
Constante	2.142248	0.00375	0.486903	0.000195	0.487079	0.000145	0.525631	0.000256

Tabla 11.9.1.1. Resultados obtenidos para el OR de los atributos en la implementación bayesiana.

	Modelo							
Atributo	*1*		*2*		*3*		*4*	
	Mínimo	Máximo	Mínimo	Máximo	Mínimo	Máximo	Mínimo	Máximo
Edad	0.947182	0.974504	0.965364	0.994035	0.966308	0.993069	0.962644	0.990758
Género femenino	0.459926	0.734094	0.464922	0.731053	0.436579	0.801209	0.484981	0.726445
Raza negra	1.39106	1.632181	N/A	N/A	0.961908	1.200851	0.832681	1.074194
puntaje COMPAS	N/A	N/A	1.257651	1.323493	1.25488	1.322819	1.23065	1.291083
Puntaje COMPAS X raza negra	N/A	N/A	N/A	N/A	N/A	N/A	1.008783	1.077364
Constante	1.968637	2.329673	0.447291	0.52968	0.451776	0.523808	0.479973	0.574359

Tabla 11.9.1.2. Rangos para el OR de los atributos en la implementación bayesiana.

	Modelo			
Medida	*1*	*2*	*3*	*4*
Chi cuadrada	141.62	736.21	757.9	709.05
LL	-3578.48	-3281.18	-3270.34	-3294.76
Pseudo-R^2	0.02	0.10	0.10	0.10

Tabla 11.9.2. Medidas de rendimiento con el método bayesiano (usando la media a posteriori).

Retomando las conclusiones de los autores, en primer lugar alegan que el OR para el término de interacción entre puntaje y raza negra, en el cuarto modelo, es evidencia de que no hay sesgo racial, pues su valor de 0.99 indicó que la variación en las posibilidades de reincidir es muy pequeña con cada aumento del término de interacción. En este caso el valor medio ascendió a 1.04, y el rango hallado fue de 1.01 a 1.08. Esto permite una conclusión más completa: aunque en promedio el valor del OR para este atributo sigue siendo cercano a la unidad, los límites reflejan que por más pequeño que sea el efecto, es seguro que hay un incremento en las posibilidades de reincidencia conforme aumenta este valor. Con cada aumento unitario para el atributo, las posibilidades aumentan entre 1% y 7%, aproximadamente, teniendo un aumento promedio de 4%.

En cuanto al cambio en las medidas de rendimiento entre los modelos 3 y 4, usando el método bayesiano, las diferencias son mucho más grandes que en el caso frecuentista. Estos resultados respaldan las conclusiones de los autores aún más que sus propios resultados, pues no sólo no mejora la capacidad predictiva al incorporar el término de interacción entre raza negra y puntaje COMPAS, sino que la misma se reduce considerablemente. Nótese que el modelo nulo, que no contiene atributos aparte de la constante *b*, requerido para calcular estas medidas, en este caso también se obtuvo con el método bayesiano. Ya que estos cambios tienen que ver con las diferencias entre coeficientes frecuentistas y las medias a posteriori bayesianas para los mismos, el cambio en el esquema de optimización entre metodologías también es una causa de los cambios en estas medidas.

Los autores también concluyeron que debido al cambio insignificante de 0.42 a 0.40 para el OR de la constante entre los modelos 2 y 3, se puede confirmar nuevamente que no hay diferencias raciales en COMPAS. Los valores medios obtenidos para este OR en estos dos modelos confirman esa conclusión, pues aquí ambos son de 0.49. La similitud también es evidente en los rangos de esta cantidad en ambos casos. Sin embargo, estos rangos también indican que en el modelo 2 el OR puede ir de 0.45 a 0.53, y en el 3 de 0.45 a 0.52. Es cierto que los cambios más extremos serían improbables, pero vale la pena mencionar que, considerando el modelo, estos serían un aumento de 0.45 a 0.52, o bien, una disminución de 0.53 a 0.45. En el primer caso, el aumento representaría que si el grupo general, representado por la constante, incluye a los elementos de raza negra, su probabilidad de reincidencia disminuye, mientras que la disminución en este OR del modelo 2 al 3 representaría un efecto opuesto. El rango para el OR del término de raza negra en el modelo 3 indica algo similar, pues tener este atributo puede representar una disminución aproximada del 4% en las posibilidades de reincidencia, pero también puede representar un aumento de hasta un 20%. El valor medio para este valor indica que en promedio se tiene un ligero aumento de 7% en las posibilidades, similar al valor de 9% hallado en la reproducción frecuentista.

Los autores concluyeron que no hay diferencias entre razas en la relación funcional entre puntaje COMPAS y reincidencia, graficando la probabilidad promedio estimada para cada puntaje COMPAS en cada raza, usando sólo el modelo 4, y obteniendo pendientes similares para ambas razas. En el caso bayesiano se puede hacer algo similar, obteniendo probabilidades predictivas con los valores medios de los coeficientes, pero también se pueden muestrear coeficientes de las distribuciones a posteriori, y obtener gráficas que reflejen la incertidumbre visualmente, como en el capítulo 7, en el ejemplo con datos sintéticos. A continuación se muestran ambos resultados.

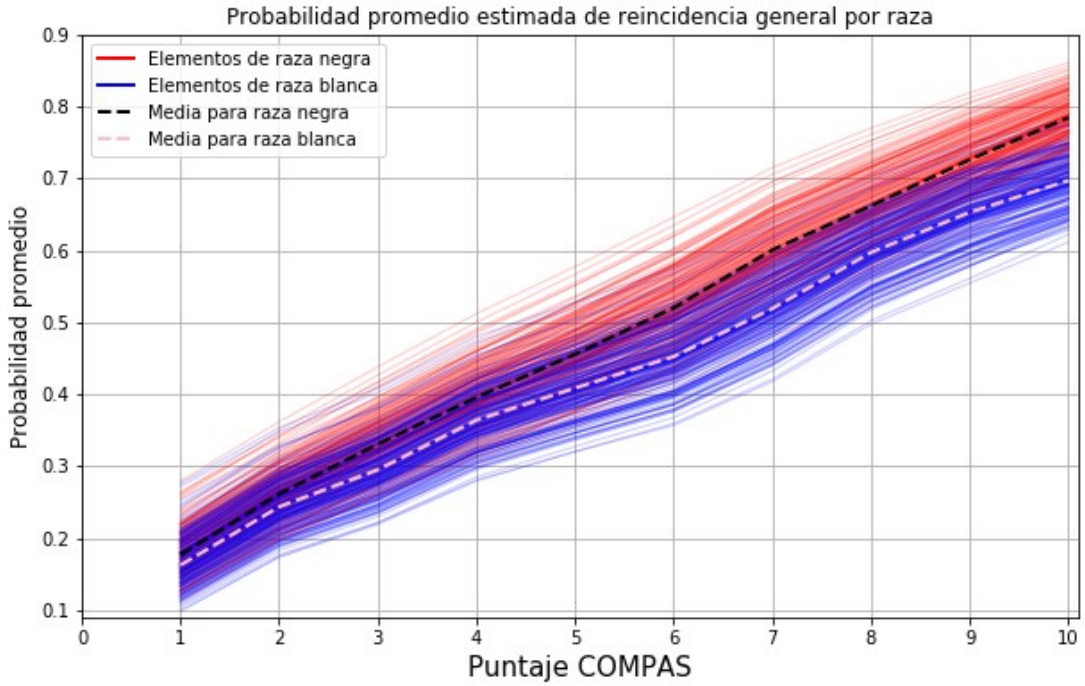


Figura 11.6.1. Relación entre puntaje COMPAS y probabilidad estimada de reincidencia, para razas negra y blanca.

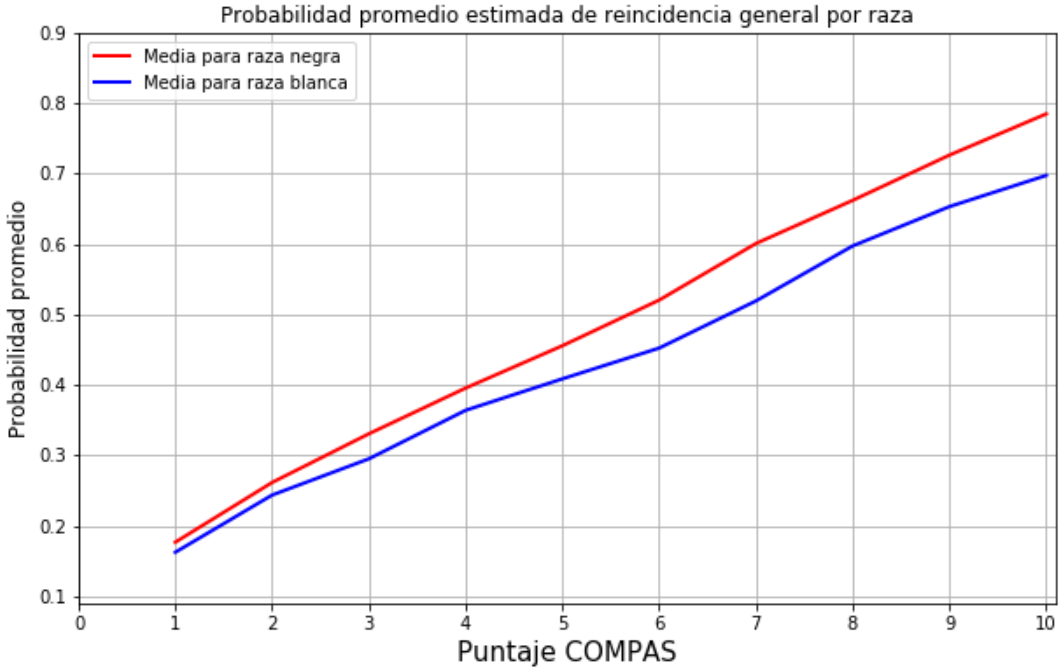


Figura 11.6.2. Relación entre puntaje COMPAS y probabilidad estimada de reincidencia, para razas negra y blanca (usando el valor medio de cada coeficiente).

En la figura 11.6.2 se puede observar que a diferencia de la gráfica obtenida en la reproducción, la cual también es similar a la presentada en el trabajo de W. Flores et al. (2016), en este caso la divergencia entre las gráficas para cada raza aumenta conforme lo hace el puntaje, mientras que en los resultados anteriores la distancia entre ambas fue más estable entre puntajes. Los resultados indican que para elementos de raza negra, la probabilidad predictiva del modelo incrementa en mayor medida con puntajes altos que para elementos de raza blanca, aunque los cambios en la pendiente son irregulares y relativamente moderados: por ejemplo, al aumentar el puntaje de 5 a 6, la diferencia entre razas es más notable que al aumentar de 9 a 10, y en el incremento de 7 a 8 parece que la probabilidad aumenta más para elementos de raza blanca, porque la pendiente para dicha raza es ligeramente mayor en ese tramo.

A su vez, la figura 11.6.1 muestra que no hay suficiente incertidumbre como para poner en duda lo que refleja la figura 11.6.2, y en general las muestras reflejan las variaciones descritas para el caso que considera sólo los valores medios. Algunas muestras para raza blanca están encima de la media para raza negra en el gráfico, y algunas de las muestras para raza negra están debajo de la media para raza blanca, pero este tipo de muestras son escasas, y en general los contornos para cada raza respaldan las conclusiones que podrían obtenerse de la media, y muestran que, considerando la incertidumbre, las diferencias son aún más moderadas: se notó que hay un aumento en la divergencia entre razas conforme aumenta el puntaje, pero la gráfica también refleja mayor incertidumbre en puntajes más altos, por lo que esta diferencia con respecto a los resultados de la reproducción y del trabajo original posiblemente deba confirmarse con más datos para ser relevante. Es útil complementar las conclusiones con una representación de la certeza de los resultados. El incremento de la dispersión en las muestras para ambas razas en puntajes más altos, sugiere que incorporar datos de individuos con puntajes en ese rango sería valioso para lograr conclusiones más certeras; esta capacidad de la metodología bayesiana, de guiar procesos de recolección de información, ya se mencionó anteriormente.

En el capítulo anterior se notó que pueden observarse las suposiciones previas con respecto a las medidas de la discriminación, calculándolas con las distribuciones a priori de los coeficientes. Ya que en este caso todos los coeficientes tienen la misma distribución normal estándar a priori, como en el capítulo anterior, los resultados para las suposiciones previas de los valores del OR, son válidas aquí, incluyendo también el caso de reincidencia violenta (porque la distribución a priori sigue siendo la misma)[88]. Tomando esos resultados, se puede decir que a priori, todos los OR para los cuatro modelos rondan entre 0.05 y 20, aproximadamente, con una media un poco mayor a 1.6. Esto quiere decir que en este caso, antes de ver los datos, se espera que para todos los atributos considerados se tenga entre un aumento de 300% y una disminución del 95% en las posibilidades de reincidir, con cada aumento unitario del atributo. En teoría, se espera que en promedio el incremento unitario de cada atributo aumente las posibilidades en poco más de 60%, pero como se mencionó en el capítulo anterior, la elevada varianza significa que este valor medio no representa un conocimiento previo firme o significativo.

En cuanto a las medidas de rendimiento y las relaciones entre puntaje y probabilidad de reincidencia (especialmente para las medidas de rendimiento), no es sensato obtener resultados a priori. En ambos casos, los resultados están basados en aplicar los modelos entrenados sobre los datos, entonces repetir los procesos usando las distribuciones a priori no sería equivalente a visualizar las suposiciones previas: lo único que se estaría haciendo es mostrar los resultados para un caso hipotético, donde no se hiciera ninguna inferencia sobre los datos, y simplemente se hicieran predicciones con base en el conocimiento previo, que en este caso en particular, es prácticamente hacer predicciones aleatorias, pues no se incorpora ningún conocimiento previo

88 Aunque en ese capítulo se presentan suposiciones a priori para tres coeficientes distintos, se hace notar que eso sólo es ilustrativo, pues en realidad los resultados de cualquier coeficiente serían válidos para cualquier otro con la misma distribución a priori.

sustancial. En otras palabras, sería probar un modelo sobre los datos, que no tuvo un proceso de aprendizaje sobre los mismos.

Aún así, se puede deducir que como las medias a priori serían cero para todos los coeficientes, los cuatro modelos tendrían rendimientos idénticos, porque todos darían la misma predicción para cualquier dato de entrada: una probabilidad de reincidencia de 0.5. Del mismo modo, esto se vería reflejado en la gráfica de la relación entre puntaje y probabilidad, pues al igual que cualquier otro atributo, el puntaje no afectaría la probabilidad predicha, y se tendría una línea horizontal fija en la probabilidad de 0.5. En cuanto a las muestras de posibles relaciones, debido a la elevada varianza a priori, se tendrían relaciones muy distintas, sin tendencias claras para la forma de la relación en el caso de cada raza.

Estos resultados no deberían interpretarse como suposiciones previas, pero puede obtenerse la gráfica de las relaciones entre puntaje y probabilidad a priori, aunque sea sólo para ilustrar lo establecido en el párrafo anterior. Nótese también que la línea fija en la probabilidad de 0.5 para la media, refleja que la reincidencia predicha es insensible al atributo de puntaje COMPAS, como lo es para todos los demás, siendo esta la razón por la cual las medidas de rendimiento serían idénticas e irrelevantes si se usan las distribuciones a priori.

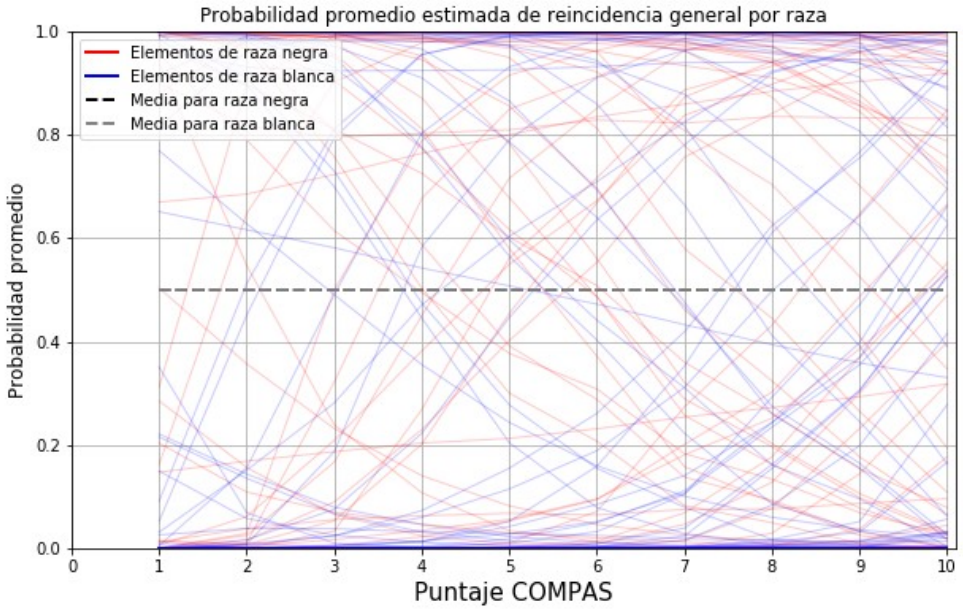


Figura 11.6.3. Relación entre puntaje COMPAS y probabilidad estimada de reincidencia para ambas razas, usando las distribuciones a priori de los coeficientes.

Los resultados a priori para el OR del capítulo anterior también son válidos en este capítulo por tener coeficientes con las mismas distribuciones a priori, por esta misma razón todo lo que se expresó aquí con respecto a las distribuciones a priori (valores del OR, medidas de rendimiento, y la relación entre puntaje COMPAS y probabilidad estimada) aplica al caso de reincidencia violenta, pues en ese caso también se tienen las mismas distribuciones a priori usadas para reincidencia general.

Reincidencia Violenta

Por último, se presentarán los resultados para el caso de reincidencia violenta, considerando datos filtrados y no filtrados, como en la reproducción. Se revisarán nuevamente las conclusiones de los autores, y se comentarán los aportes de la metodología bayesiana en esa parte del proceso.

Coeficientes de Regresión Logística (componentes del vector **w** y valor de *b*)										
	Distribuciones a priori		*Distribuciones a posteriori*							
	Modelo		Modelo							
Atributo correspondiente en el vector **x**	*Todos los Modelos (1-4)*		*Modelo 1*		*Modelo 2*		*Modelo 3*		*Modelo 4*	
	Media	Varianza	Media	Varianza	Media	Varianza	Media	Varianza	Media	Varianza
Ninguno (el coeficiente es la constante *b*)	0.0	1.0	-0.285242	0.002109	-1.57935	0.002461	-1.58138	0.002946	-1.33838	0.002112
Edad (A*ge*)	0.0	1.0	-0.060125	3.3e-05	-0.046891	3.4e-05	-0.04863	3.2e-05	-0.043856	3.1e-05
Género femenino (F*emale*)	0.0	1.0	-0.839435	0.020095	-0.852763	0.019926	-0.860053	0.014671	-0.806599	0.01242
Raza negra (*Black*)	0.0	1.0	0.46878	0.002273	N/A	N/A	0.062118	0.002992	-0.505066	0.003656
puntaje COMPAS (*NPC Decile*)	0.0	1.0	N/A	N/A	0.250691	0.000103	0.245008	7.4e-05	0.161901	0.000117
Puntaje COMPAS X raza negra (*NPC Decile X black*)	0.0	1.0	N/A	N/A	N/A	N/A	N/A	N/A	0.104427	0.000134

Tabla 11.10.1. Distribuciones de los coeficientes en la implementación bayesiana para reincidencia violenta (datos filtrados: 3377 registros).

Coeficientes de Regresión Logística (componentes del vector **w** y valor de *b*)										
	Distribuciones a priori		*Distribuciones a posteriori*							
	Modelo		Modelo							
Atributo correspondiente en el vector **x**	*Todos los Modelos (1-4)*		*Modelo 1*		*Modelo 2*		*Modelo 3*		*Modelo 4*	
	Media	Varianza	Media	Varianza	Media	Varianza	Media	Varianza	Media	Varianza
Ninguno (el coeficiente es la constante *b*)	0.0	1.0	-0.205069	0.001692	-1.40498	0.002065	-1.392897	0.001978	-1.158307	0.002757
Edad (A*ge*)	0.0	1.0	-0.033097	3.2e-05	-0.02565	3.3e-05	-0.025791	2.8e-05	-0.028106	2.8e-05
Género femenino (F*emale*)	0.0	1.0	-0.751698	0.013997	-0.785922	0.014376	-0.789776	0.014769	-0.750599	0.014142
Raza negra (*Black*)	0.0	1.0	0.456074	0.002842	N/A	N/A	0.073839	0.002581	-0.482526	0.001636
puntaje COMPAS (*NPC Decile*)	0.0	1.0	N/A	N/A	0.25953	0.000153	0.25293	0.0001	0.173505	8.5e-05
Puntaje COMPAS X raza negra (*NPC Decile X black*)	0.0	1.0	N/A	N/A	N/A	N/A	N/A	N/A	0.139878	0.000142

Tabla 11.10.2. Distribuciones de los coeficientes en la implementación bayesiana para reincidencia violenta (datos sin filtrar: 3967 registros).

Como en el caso frecuentista, las diferencias entre coeficientes, que dependen de los datos usados, persisten al usar la metodología bayesiana. En general las medias obtenidas y las varianzas presentan cambios, aunque no son extremos, dependiendo del filtrado. También hay cambios en general entre las medias reportadas y los valores puntuales obtenidos para sus contrapartes frecuentistas que, como ya se ha mencionado, es atribuible al cambio en la optimización entre un método y otro, y no sólo al cambio de metodología.

A continuación, se muestran gráficamente las distribuciones a posteriori para los coeficientes, usando los datos filtrados y sin filtrar. Como siempre, se tiene la distribución a priori común como referencia. Para facilitar las comparaciones entre el uso de datos filtrados y sin filtrar, los coeficientes, para cada uno de los cuatro modelos, se presentan para ambos casos en una misma figura.

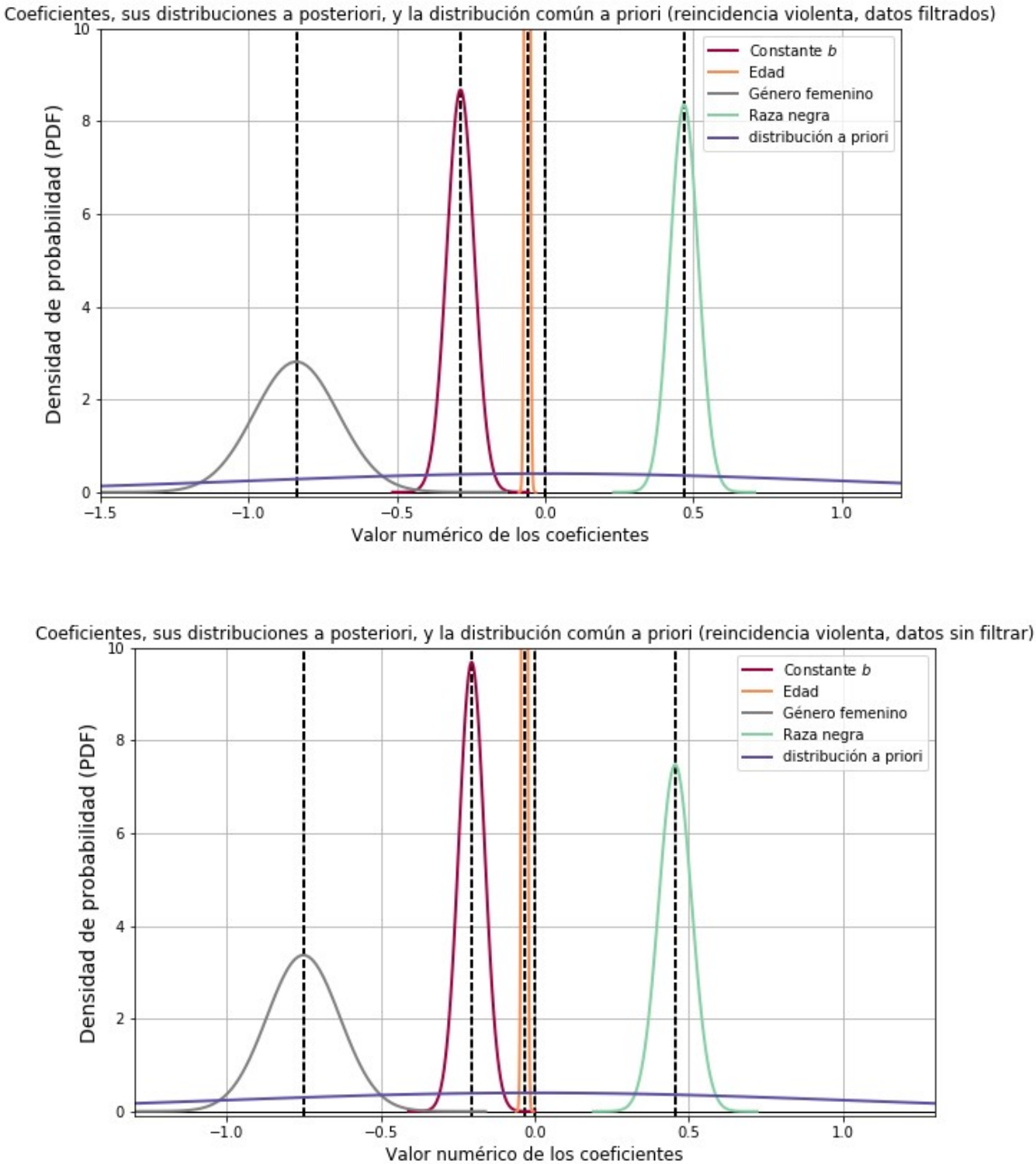
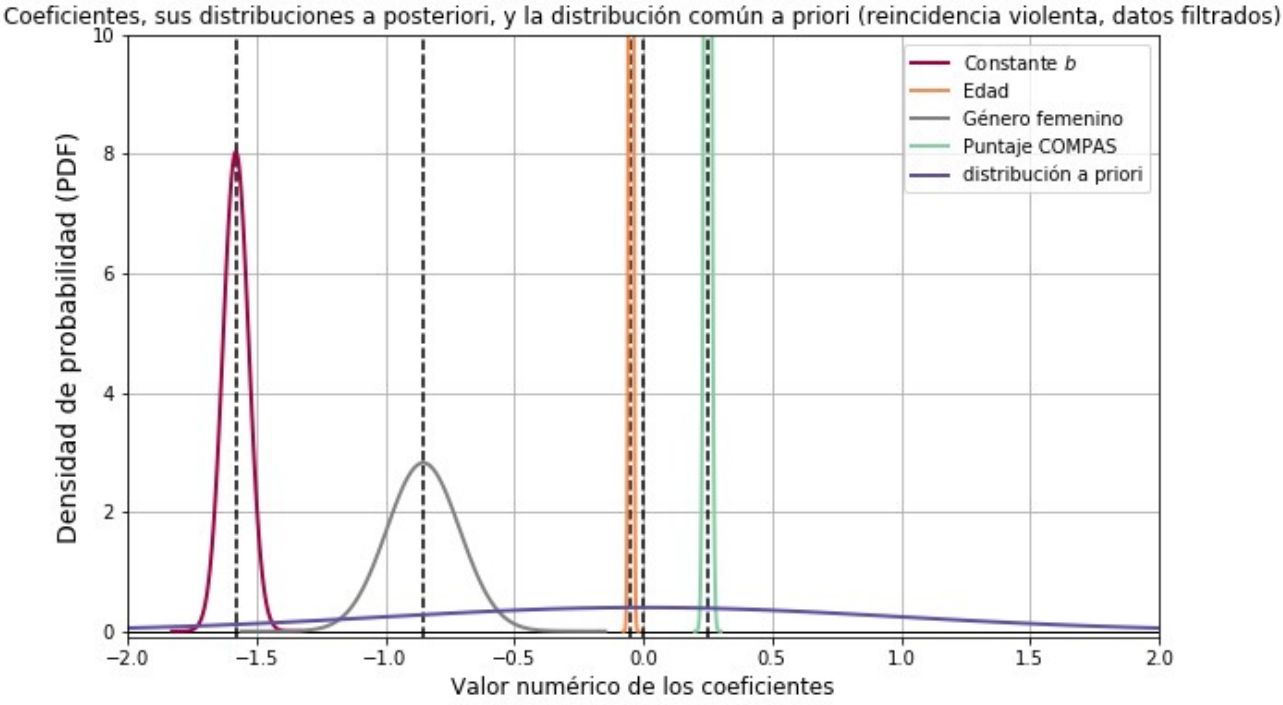


Figura 11.7.1. Distribuciones para los coeficientes del modelo 1 de regresión logística bayesiana (reincidencia violenta), para los datos filtrados de 3377 registros (arriba) y los datos no filtrados de 3967 registros (abajo).



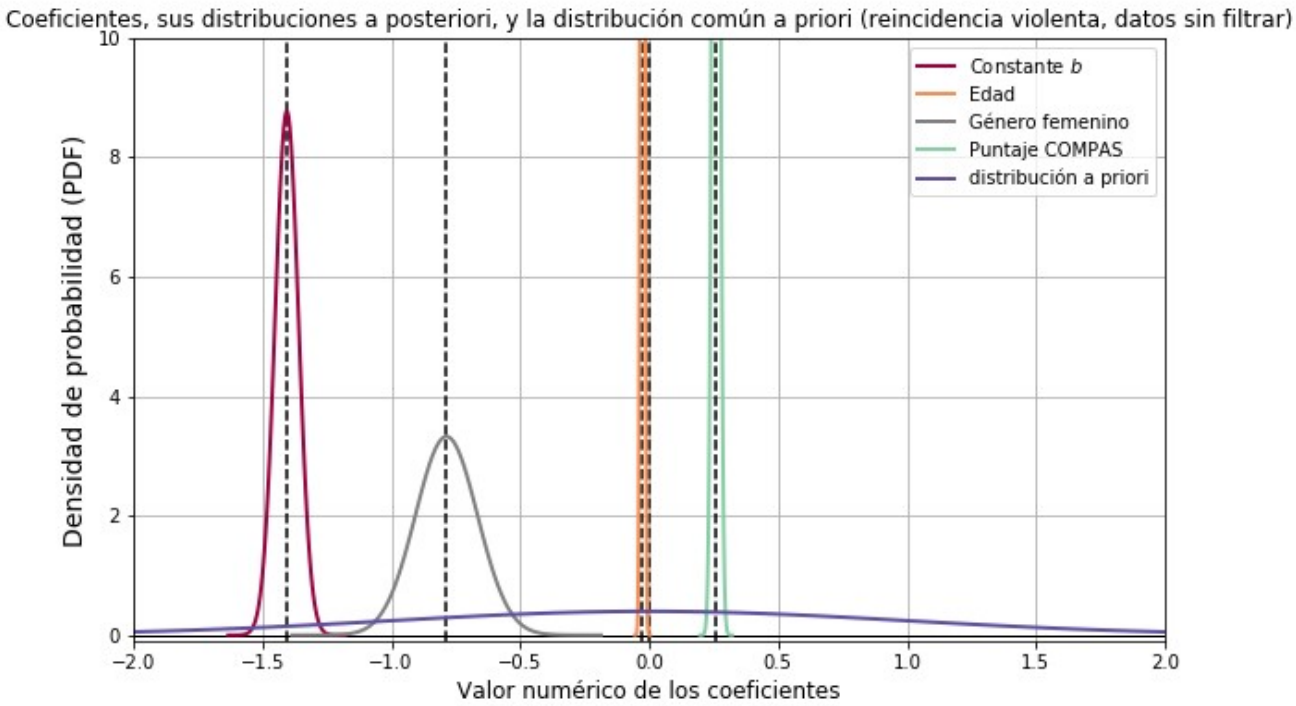


Figura 11.7.2. Distribuciones para los coeficientes del modelo 2 de regresión logística bayesiana (reincidencia violenta), para los datos filtrados de 3377 registros (arriba) y los datos no filtrados de 3967 registros (abajo).

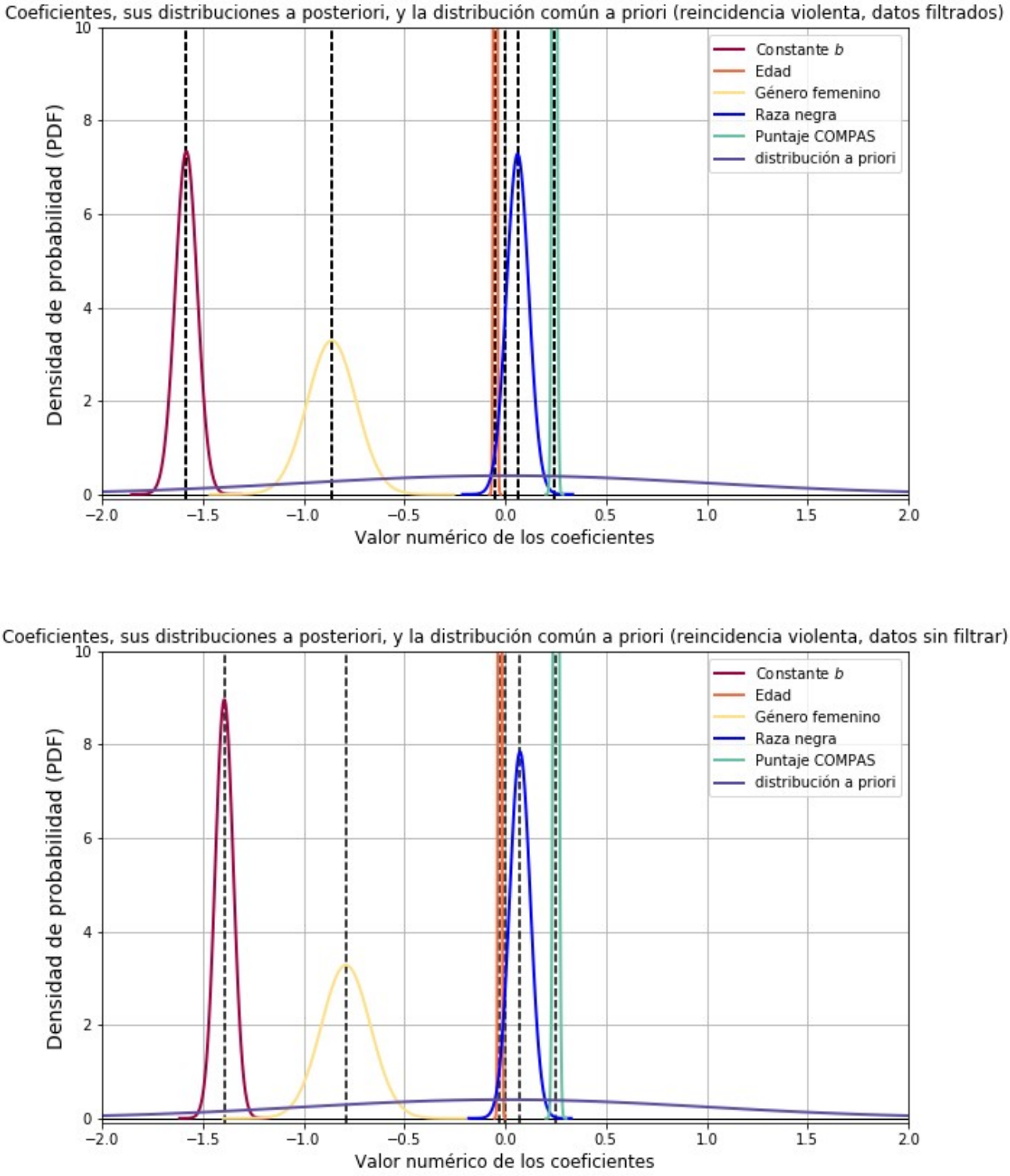


Figura 11.7.3. Distribuciones para los coeficientes del modelo 3 de regresión logística bayesiana (reincidencia violenta), para los datos filtrados de 3377 registros (arriba) y los datos no filtrados de 3967 registros (abajo).

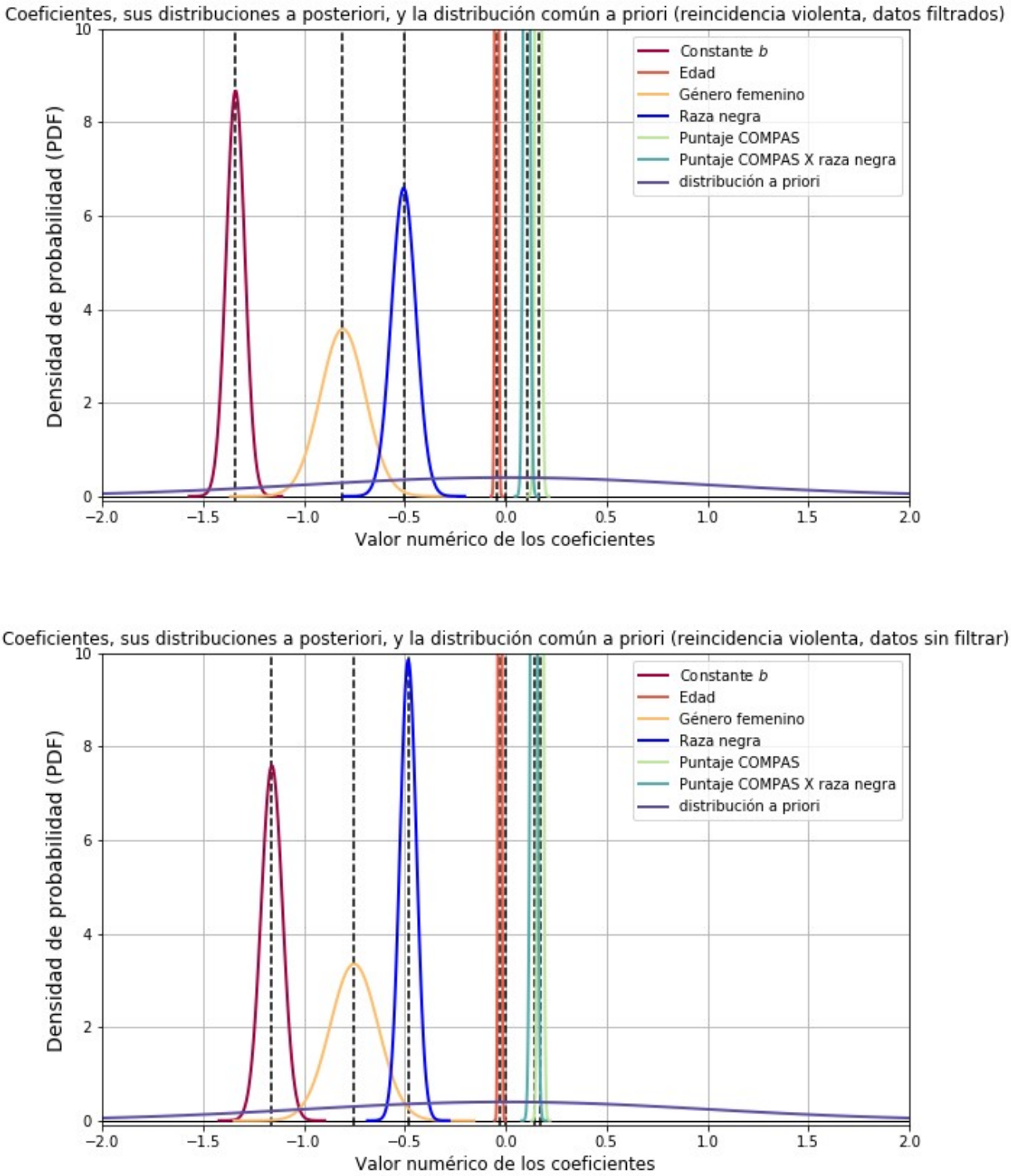


Figura 11.7.4. Distribuciones para los coeficientes del modelo 4 de regresión logística bayesiana (reincidencia violenta), para los datos filtrados de 3377 registros (arriba) y los datos no filtrados de 3967 registros (abajo).

En este caso, los coeficientes para reincidencia violenta muestran varias características parecidas a los coeficientes para reincidencia general. Claramente, los valores medios y varianzas para estos coeficientes son diferentes, pero los cambios que cada coeficiente experimenta de un modelo a otro son similares. Nuevamente se puede apreciar que ningún coeficiente tiene tanta incertidumbre como se llegó a apreciar en el caso de ProPublica, y el coeficiente de género femenino es nuevamente el que mayor incertidumbre tiene (aunque nuevamente su media es más estable entre modelos que las de otros coeficientes con menos varianza).

Algo interesante es que al ilustrar la incertidumbre sobre los coeficientes, se puede notar cómo, para cada modelo, las diferencias al filtrar o no los datos parecen ser aún menos relevantes. Para cada modelo, la cercanía de las medias y el valor de la varianza para cada coeficiente permite que el valor medio encontrado para los datos filtrados sea un valor plausible para la distribución a posteriori aprendida usando los datos sin filtrar, y viceversa. En el caso frecuentista, cuando se obtuvieron los OR se determinó que estas diferencias no causaron cambios significativos en las conclusiones, e incluso los cambios numéricos de los coeficientes se consideraron moderados a simple vista. Sin embargo, el método bayesiano permite visualizar con claridad cuán plausibles son los resultados de cada caso (filtrar o no los datos) para el otro en cada modelo, y esta ventaja se extiende a otras cantidades derivadas de los coeficientes y sus distribuciones a posteriori. Finalmente, la distribución a priori nuevamente fue suficientemente amplia para aprender los valores de todos los coeficientes de forma adecuada.

	Modelo							
Atributo	*1*		*2*		*3*		*4*	
	Media	Varianza	Media	Varianza	Media	Varianza	Media	Varianza
Edad	0.941615	2.9e-05	0.95416	3.1e-05	0.952557	2.9e-05	0.957096	2.9e-05
Género femenino	0.436085	0.003827	0.430276	0.003694	0.426095	0.002697	0.449289	0.002527
Raza negra	1.599551	0.005798	N/A	N/A	1.065712	0.003387	0.60498	0.001342
puntaje COMPAS	N/A	N/A	1.284927	0.00017	1.277655	0.00012	1.175819	0.000161
Puntaje COMPAS X raza negra	N/A	N/A	N/A	N/A	N/A	N/A	1.110195	0.000165
Constante	0.752664	0.001187	0.206374	0.000104	0.206023	0.000125	0.262585	0.000145

Tabla 11.11.1.1. Resultados obtenidos para el OR de los atributos en la implementación bayesiana para reincidencia violenta (datos filtrados: 3377 registros).

	Modelo							
Atributo	*1*		*2*		*3*		*4*	
	Media	Varianza	Media	Varianza	Media	Varianza	Media	Varianza
Edad	0.967413	3e-05	0.974644	3.2e-05	0.974562	2.7e-05	0.97229	2.7e-05
Género femenino	0.474672	0.003148	0.458786	0.003021	0.457138	0.003125	0.475586	0.003227
Raza negra	1.579769	0.007073	N/A	N/A	1.078053	0.00299	0.618007	0.000626
puntaje COMPAS	N/A	N/A	1.296355	0.000255	1.28783	0.000164	1.189522	0.00012
Puntaje COMPAS X raza negra	N/A	N/A	N/A	N/A	N/A	N/A	1.150265	0.000188
Constante	0.815319	0.001117	0.245638	0.000124	0.248629	0.000123	0.314495	0.000271

Tabla 11.11.1.2. Resultados obtenidos para el OR de los atributos en la implementación bayesiana para reincidencia violenta (datos sin filtrar: 3967 registros).

	Modelo							
Atributo	*1*		*2*		*3*		*4*	
	Mínimo	Máximo	Mínimo	Máximo	Mínimo	Máximo	Mínimo	Máximo
Edad	0.926086	0.957832	0.938206	0.970822	0.93656	0.968736	0.941418	0.973113
Género femenino	0.284986	0.652874	0.281707	0.643109	0.297597	0.608734	0.320322	0.620977
Raza negra	1.391295	1.840197	N/A	N/A	0.907176	1.249056	0.503578	0.720958
puntaje COMPAS	N/A	N/A	1.247501	1.324166	1.245414	1.31159	1.139218	1.214436
Puntaje COMPAS X raza negra	N/A	N/A	N/A	N/A	N/A	N/A	1.073058	1.148595
Constante	0.657178	0.860462	0.178226	0.238457	0.174445	0.24157	0.228911	0.29998

Tabla 11.11.2.1. Rangos para el OR de los atributos en la implementación bayesiana para reincidencia violenta (datos filtrados: 3377 registros).

	Modelo							
Atributo	*1*		*2*		*3*		*4*	
	Mínimo	Máximo	Mínimo	Máximo	Mínimo	Máximo	Mínimo	Máximo
Edad	0.95173	0.983785	0.958489	0.991514	0.959284	0.989998	0.957188	0.987704
Género femenino	0.333275	0.665673	0.320565	0.65	0.31889	0.653843	0.331314	0.671452
Raza negra	1.351419	1.847517	N/A	N/A	0.928357	1.249442	0.546849	0.695226
puntaje COMPAS	N/A	N/A	1.250605	1.344594	1.250202	1.327582	1.157945	1.222698
Puntaje COMPAS X raza negra	N/A	N/A	N/A	N/A	N/A	N/A	1.110609	1.19131
Constante	0.722083	0.919279	0.214781	0.280431	0.216994	0.283323	0.268811	0.366121

Tabla 11.11.2.2. Rangos para el OR de los atributos en la implementación bayesiana para reincidencia violenta (datos sin filtrar: 3967 registros).

	Datos Filtrados (N = 3377)				*Datos sin Filtrar (N = 3967)*			
	Modelo				*Modelo*			
Medida	*1*	*2*	*3*	*4*	*1*	*2*	*3*	*4*
Chi cuadrada	23.18	268.38	248.53	250.19	96.33	318.90	310.14	338.75
LL	-1534.15	-1411.56	-1421.48	-1420.65	-1770.10	-1658.82	-1663.20	-1648.89
Pseudo-R^2	0.01	0.09	0.08	0.08	0.03	0.09	0.09	0.09

Tabla 11.11.3. Medidas de rendimiento con el método bayesiano para reincidencia violenta, usando la media a posteriori, para el caso con datos filtrados y sin filtrar.

Los resultados para los OR en el caso bayesiano tienen diferencias importantes con respecto al caso frecuentista, y por ende, con respecto a las conclusiones sugeridas por los resultados de la reproducción y las que presentan los autores. Como podría esperarse, considerando las similitudes en las distribuciones a posteriori, no hay cambios significativos en los OR para datos filtrados y sin filtrar, por lo que ambos llevan a las mismas conclusiones y diferencias ante el método frecuentista para esta medida. Sin embargo, hay cambios importantes en las medidas de rendimiento calculadas, no sólo entre datos filtrados y no filtrados, sino también entre cada caso y su contraparte frecuentista.

En primer lugar, estos resultados muestran que el OR para el término de interacción entre puntaje COMPAS y raza negra en el modelo 4, sí denota un aumento claro en las posibilidades de reincidencia, contrario a la conclusión de los autores. En el caso frecuentista se hallaron valores de 1.03 y 1.04 para datos filtrados y sin filtrar, respectivamente, que de por sí eran mayores al valor de 1.01 presentado por los autores. En este caso los valores ascienden a medias de 1.11 y 1.15, respectivamente, y además se determinaron rangos de 1.07 a 1.15 para datos filtrados, y de 1.11 a 1.19 para datos sin filtrar. Esto indica que no sólo se encuentra un efecto mayor, sino que también se puede afirmar con certeza[89].

En cuanto a las diferencias en el rendimiento entre modelos 3 y 4, para datos filtrados, al igual que en todos los resultados de la reproducción (incluyendo los de datos no filtrados), se confirma que el cambio es mínimo (aunque se puede considerar una mejora a fin de cuentas). Sin embargo, los resultados para datos no filtrados indicaron una mejora que no es drástica, pero tampoco es despreciable como en la mayoría de los casos anteriores. Esto no es tan raro, ya que el OR para el término de interacción del modelo 4 para datos sin filtrar también fue el más elevado de entre todas las implementaciones (bayesianas, frecuentistas, de reincidencia general o violenta, y con datos filtrados o sin filtrar).

En cuanto a la afirmación de que el atributo de raza negra es insignificante cuando está presente el atributo de puntaje COMPAS, para el modelo 3 los valores medios bayesianos para el atributo de raza respaldan dicha conclusión más que los resultados de la reproducción y de los autores. El rango indica que la relación podría indicar incluso menor probabilidad de reincidencia para elementos de raza negra, pues el límite inferior es menor a la unidad para datos filtrados y sin filtrar, y a su vez, los límites superiores son apenas mayores que el valor de 1.24 de los autores, que como se mencionó antes, parece precipitado considerar insignificante. Para el modelo 4 esto es aún más marcado, pues los límites superiores para el OR del atributo son 0.72 y 0.70, indicando con gran certeza que hay menor probabilidad para elementos de raza negra, aunque debe considerarse que en este modelo, el hecho de que el término de interacción entre puntaje COMPAS y raza negra indique, también con gran certeza, un aumento en las probabilidades, es una objeción a concluir que los elementos de raza negra tengan una ventaja en este modelo. Desde la reproducción, esta conclusión de los autores se ponía en duda, y aunque los resultados bayesianos para el modelo 3 favorezcan más su conclusión que sus propios resultados, aún no es suficiente para validar la misma (en especial por lo observado en el modelo 4).

Finalmente, los autores también sustentaron su conclusión en el hecho de que el OR para el término constante no varía entre los modelos 2 y 3. Como en el caso de reincidencia general, esto se confirma con el mínimo cambio registrado entre las medias y rangos para este valor, tanto en datos filtrados como en datos sin filtrar[90]. Como ya se mencionó en la implementación bayesiana de reincidencia general, los rangos permiten contemplar las posibilidades en los extremos, que aunque son improbables, permiten decir que aunque se esperan cambios mínimos en este OR al pasar del modelo 2 a 3, aún es posible observar cambios de hasta 0.07 en el valor total, según los datos obtenidos de las distribuciones a posteriori del modelo.

89 Es importante recalcar que en este trabajo lo importante son las conclusiones que pueden obtenerse de los resultados bayesianos, no precisamente contradecir los trabajos originales. Para poder afirmar qué modelo tiene respuestas más acertadas, se deben realizar comparaciones de rendimiento. Podrían compararse, entre métodos, las medidas de rendimiento obtenidas para los modelos del 1 al 4, o bien, como en el anexo, usar medidas que reflejen la capacidad de generalización de los modelos implementados con cada método. En cualquier caso, el objetivo en el escrito principal es más bien mostrar cómo las respuestas bayesianas pueden ser más completas, informativas e intuitivas.

90 Esto se refiere a los cambios entre los modelos 2 y 3 para datos filtrados, y entre los modelos 2 y 3 para datos sin filtrar, no a los cambios al filtrar o no los datos.

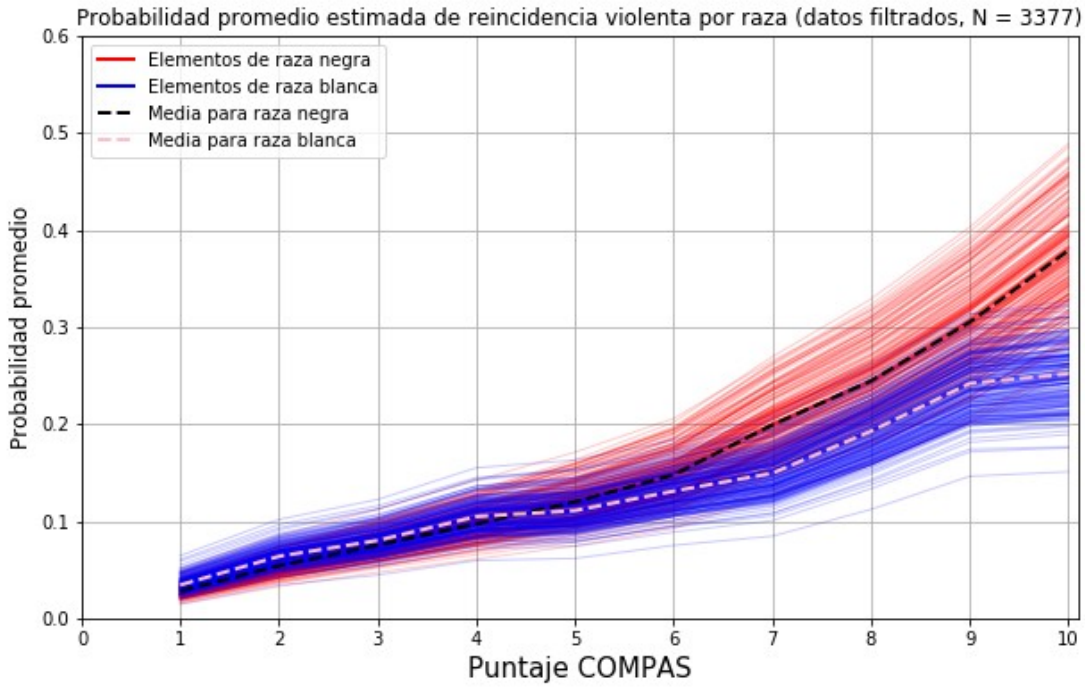


Figura 11.8.1.1. Relación entre puntaje COMPAS y probabilidad estimada de reincidencia violenta, para razas negra y blanca. Caso para datos filtrados.

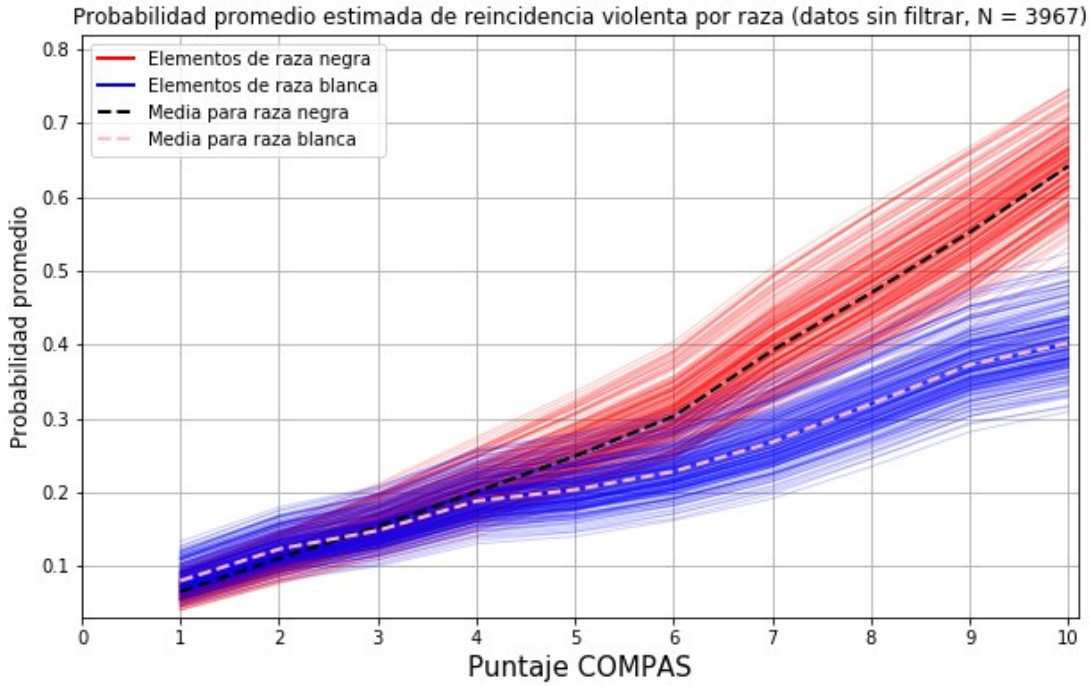


Figura 11.8.1.2. Relación entre puntaje COMPAS y probabilidad estimada de reincidencia violenta, para razas negra y blanca. Caso para datos sin filtrar.

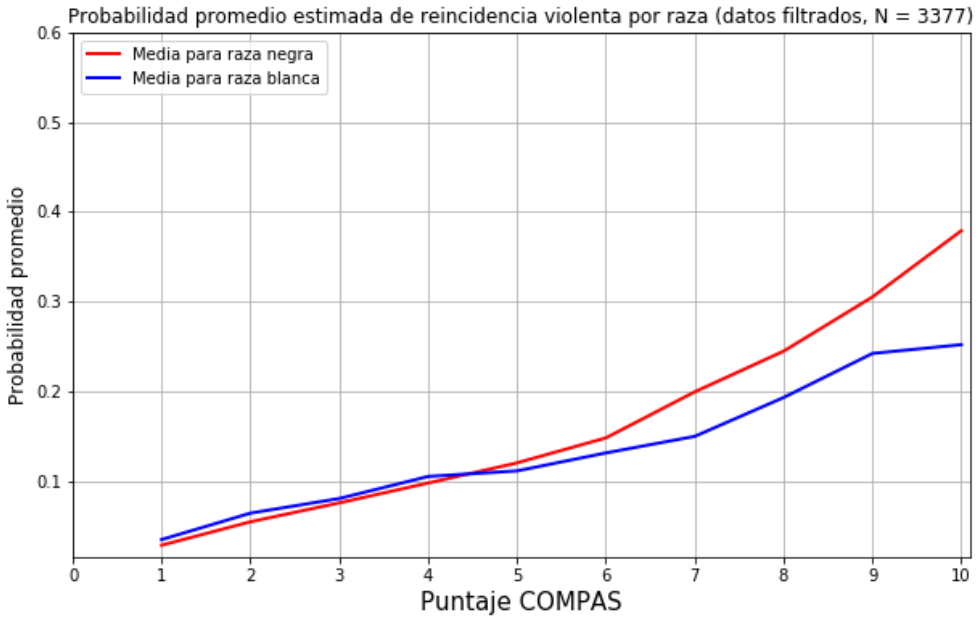


Figura 11.8.2.1. Relación entre puntaje COMPAS y probabilidad estimada de reincidencia violenta, para razas negra y blanca. Caso para datos filtrados, usando el valor medio de cada coeficiente.

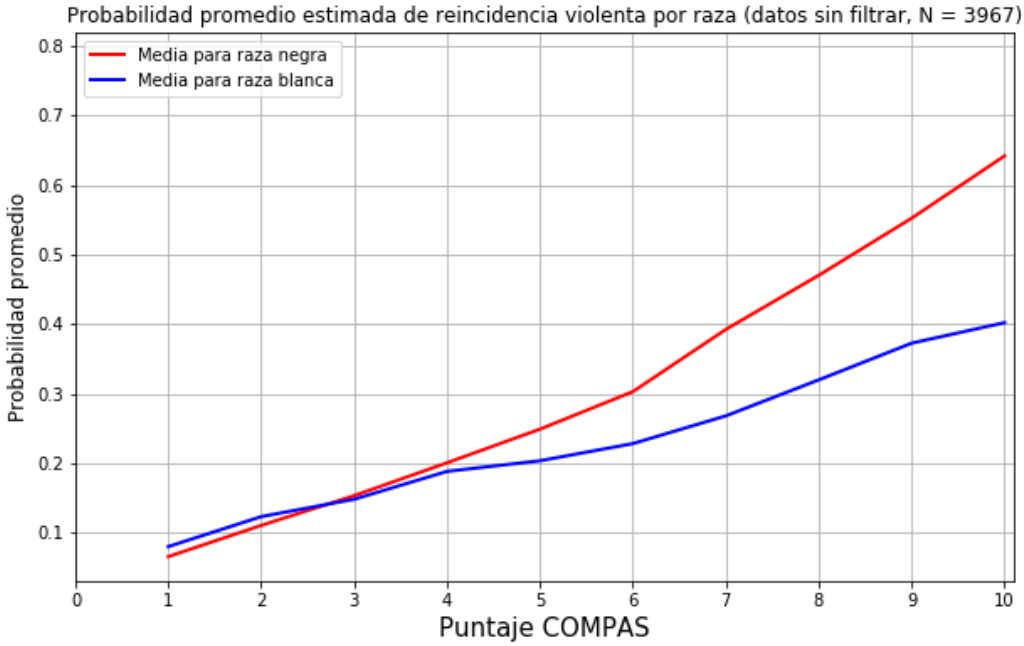


Figura 11.8.2.2. Relación entre puntaje COMPAS y probabilidad estimada de reincidencia violenta, para razas negra y blanca. Caso para datos sin filtrar, usando el valor medio de cada coeficiente.

De las figuras 11.8.2.1 y 11.8.2.2, que se generan usando los valores medios a posteriori para los coeficientes, se confirma lo visto en la reproducción, donde se notó una divergencia entre razas que aumenta con el puntaje. En este caso la divergencia fue mucho más pronunciada para el caso con datos sin filtrar, mientras que para los datos filtrados el resultado no es tan diferente del obtenido en la reproducción frecuentista. Algo interesante es que en esta implementación, tanto para datos filtrados y datos sin filtrar, la probabilidad predicha para raza blanca inicia siendo ligeramente superior, mientras que en el caso frecuentista en ningún punto superó a la probabilidad predictiva para la raza negra.

Considerando ahora las gráficas generadas con muestras de las distribuciones a posteriori, la divergencia en puntajes altos es evidente, y la separación es notable comparando con el caso de reincidencia general. Nuevamente se observa una mayor dispersión en puntajes altos, indicando que adquirir datos de individuos en ese rango llevará a un modelo con mayor certeza en las predicciones. El efecto de tener más datos se puede notar en la diferencia de la gráfica para datos no filtrados y datos sin filtrar, donde, especialmente para puntajes altos, la dispersión para las predicciones del modelo de datos sin filtrar, es considerablemente menor: esto era de esperarse, porque el modelo que usó datos sin filtrar, incluyó 3967 registros en su aprendizaje, mientras que el modelo con datos filtrados usó sólo 3377. Finalmente, hay que recordar que las probabilidades en estas gráficas se estiman usando el modelo 4.

REFERENCIAS

Allison, P. D. (s/f). *Measures of Fit for Logistic Regression*. 13.

Glm.fit function | R Documentation. (s/f). Recuperado el 16 de Junio de 2020, de

https://www.rdocumentation.org/packages/scidb/versions/1.2-0/topics/glm.fit

Julia Angwin, J. L. (2016, mayo 23). *Machine Bias* [Text/html]. ProPublica.

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Lectur21. (s/f). Recuperado el 29 de junio de 2020, de

http://web.pdx.edu/~newsomj/pa551/lectur21.htm

MLPR w6c—Machine Learning and Pattern Recognition. (s/f). Recuperado el 29 de junio de 2020, de

https://www.inf.ed.ac.uk/teaching/courses/mlpr/2018/notes/w6c_bayesian_inference_prediction.html

Simple Logistic Regression. (s/f). Recuperado el 23 de junio de 2020, de

http://vassarstats.net/logreg1.html

Tfp.glm.fit | TensorFlow Probability. (s/f). TensorFlow. Recuperado el 15 de junio de 2020, de

https://www.tensorflow.org/probability/api_docs/python/tfp/glm/fit

W. Flores, A., Bechtel, K., & Lowenkamp, C. (2016). False Positives, False Negatives, and False Analyses: A Rejoinder to “Machine Bias: There’s Software Used Across the Country to Predict Future Criminals. And it’s Biased Against Blacks.” *Federal probation*, *80*.

12. CONCLUSIONES

En los dos estudios presentados en este trabajo, se exploraron implicaciones importantes de usar la metodología bayesiana en el estudio del sesgo. Las conclusiones derivadas de los modelos implementados no muestran cambios significativos atribuibles a la metodología bayesiana en el más alto nivel: las diferencias de este tipo se dieron sólo en el segundo análisis, pero debido a las inconsistencias en los datos usados por los autores para estudiar la reincidencia violenta. Esto sólo significa que las tareas desarrolladas son, como muchos otros problemas simples, casos donde ambos métodos finalmente llegan a la misma respuesta. Las diferencias al usar el método bayesiano son del tipo que se esperaba desde un inicio, y que se expuso al ilustrar la metodología con el ejemplo sobre datos sintéticos en el capítulo 7: se derivan conclusiones más completas, que consideran la incertidumbre sobre los resultados obtenidos y son fácilmente interpretables; aún si a fin de cuentas ambos métodos, a grandes rasgos, llegan a la misma conclusión.

Como se ha reiterado, este trabajo es superficial en muchos aspectos. Las tareas resueltas y los modelos implementados palidecen en complejidad comparados al estado del arte. Además, la naturaleza de la metodología bayesiana, y el debate entre la misma y la metodología frecuentista, son temas explorados a un grado superior al de este trabajo, el cual es un simple ejemplo más de lo ya establecido. En ningún momento se esperó tener resultados innovadores en relación con esta discusión.

A pesar de esto, el trabajo cumplió exitosamente los objetivos, pues el punto era ilustrar las cualidades principales de esta metodología, bien establecida, en términos de su aplicación en el estudio del sesgo. Respondiendo a la pregunta de investigación planteada, las implicaciones de usar la metodología bayesiana en este caso son las mismas que en otras aplicaciones: deben hacerse consideraciones adicionales que implican diseños e implementaciones más laboriosos, pero a cambio se tienen respuestas más profundas e intuitivas. En el caso del sesgo, y en aplicaciones tan sensibles como la predicción de la reincidencia delictiva, esto es más que una lista de ventajas para quien toma las decisiones: en estas aplicaciones, tener resultados interpretables, explícitos con respecto a su certeza y presunciones, es una ganancia que va más allá de la precisión, la eficacia o el rendimiento.

Al derivar conclusiones con la metodología bayesiana, se pudieron discutir con mayor profundidad los resultados en cada estudio que al usar el método ortodoxo. En el caso de ProPublica, las conclusiones de los autores están basadas en la medida de riesgo relativo (RR). La metodología bayesiana permitió conclusiones más completas, incorporando la incertidumbre sobre esta medida y también la noción del valor a priori como la expresión de suposiciones previas. Además, la incertidumbre sobre los coeficientes fue indicativa de qué tipos de datos se requieren en mayores cantidades para tener más certeza en las conclusiones que los involucren.

En el caso de la réplica, se observó lo mismo para la medida del radio de posibilidades (OR) y las conclusiones derivadas de la misma, pero el método bayesiano también se usó para profundizar el estudio de la relación entre puntaje COMPAS y la probabilidad de reincidencia predicha por el modelo para cada raza, incorporando y visualizando la incertidumbre para la relación. La incertidumbre sobre los coeficientes y en las gráficas de las relaciones entre puntaje y probabilidad, nuevamente fue una guía potencial para la recolección de datos a futuro, y además, en el caso de reincidencia violenta, mostró el impacto de filtrar o no los datos en las conclusiones obtenidas, con mayor profundidad que las diferencias puntuales del método frecuentista. En las medidas de rendimiento, no hay diferencias relevantes en el tipo de resultados obtenidos, pues simplemente se pasa de usar los valores puntuales frecuentistas a usar los valores de las medias a posteriori, que también son puntuales. Al representar suposiciones con distribuciones a priori, para el OR el análisis se desarrolló de la misma manera que en el caso de ProPublica. También se revisó la obtención de las medidas de rendimiento y de las relaciones entre puntaje y probabilidad predicha, usando distribuciones a priori, puntualizando que no deben interpretarse como

suposiciones previas, pero sí son resultados que se pueden obtener fácilmente, y permiten explorar otros aspectos de los modelos para esta aplicación.

Ambos casos de estudio reflejaron que las conclusiones frecuentistas y bayesianas son, en esencia, las mismas, al menos para estos problemas, y lo valioso del método bayesiano es la profundidad añadida a las conclusiones. En términos más literales: aunque el valor puntual frecuentista sea similar a la media a posteriori del método bayesiano, que se puede considerar la mejor respuesta de este último cuando es forzado a dar una solución puntal, la ventaja al momento de tomar decisiones y derivar conclusiones es que el método bayesiano, además de la media, provee una distribución a posteriori cuya forma y varianza permite emitir juicios más complejos e informados.

Trabajo a Futuro

Hay mucho más que podría desarrollarse sobre la base de este trabajo, y parte de ello se ha introducido en el anexo de este trabajo. Varias secciones del anexo plantean nuevas direcciones potenciales para investigaciones a futuro, como el estudio más riguroso del rendimiento de los modelos, o realizar pruebas más extensas con los procesos de aprendizaje y optimización, o estudiar más a fondo alternativas para las distribuciones a priori.

En cuanto al área de aplicación, este trabajo está enfocado en la evaluación de la herramienta COMPAS y no en la tarea de clasificación que resuelve la herramienta misma, que es una perspectiva más común al abordar el problema del sesgo en el aprendizaje computacional. Aunque hasta cierto punto las bases para esto se presentaron en el ejemplo con datos sintéticos, podría hacerse un estudio enfocado específicamente a COMPAS y la predicción de la reincidencia. Por ejemplo, si la tarea que COMPAS resuelve se llevara a cabo con un método bayesiano, en vez de asignar un puntaje del 1 al 10 para cada individuo, podría asignarse una distribución, indicando un puntaje medio con un rango de valores que refleje la incertidumbre para el puntaje otorgado. Individuos con el mismo puntaje medio podrían tener grados de certeza muy distintos en el resultado, y esto proveería información valiosa para quien usa el puntaje al tomar decisiones. Un estudio de este tipo implicaría investigar el funcionamiento interno de COMPAS y analizar con más detalle cómo se aplican los puntajes en casos reales.

Aún hay más posibilidades, como tratar con modelos más complejos que la regresión logística, o incorporar nociones de justicia en el algoritmo de aprendizaje, ya sea en la tarea de descubrimiento o prevención del sesgo, pero en la mayoría de los casos no deberían considerarse extensiones de este trabajo, sino estudios independientes con enfoques más precisos y complejos. El enfoque de este trabajo es de un nivel general y rudimentario, que permite considerar la enorme cantidad de comparaciones posibles entre modelos frecuentistas y bayesianos como extensiones posibles del mismo. Sin embargo, la literatura actual referente a estas comparaciones en la aplicación específica del prejuicio, sesgo, o justicia en el aprendizaje computacional, no es extensa, y suele ir en direcciones que son, en concepto, distintas. Considerando esto, no parece tan descabellado que las extensiones posibles al trabajo sean tan numerosas.

ANEXO: RESULTADOS ADICIONALES

En esta sección se presentan distintos resultados que se mantuvieron fuera del cuerpo principal del escrito para no entorpecer la discusión principal, pero que de cualquier modo es importante tomar en cuenta. Varios de estos resultados se comentaron en el escrito, refiriendo al lector a esta sección para análisis más profundos o evidencias de distintas afirmaciones planteadas.

Al igual que para el resto del escrito, el código para esta sección puede consultarse en la siguiente liga: https://github.com/gerkbyrd/tesis_sesgo_ML_clasificadores_bayesianos.

DISTRIBUCIONES DE PUNTAJE COMPAS POR RAZA

Como se mencionó en el cuerpo del escrito, una parte de la exploración de datos que se omitió en la reproducción, pero estaba presente en el trabajo original de ProPublica (*propublica/compas-analysis*, 2016/2020) fue mostrar la distribución de puntajes COMPAS para cada raza, tanto en el caso de puntajes de reincidencia general, como en el de puntajes de reincidencia violenta. A continuación se presenta dicha información. Nótese que la información para razas afroamericana y caucásica es la misma que se presentó en los gráficos dentro del capítulo correspondiente a la reproducción de este análisis.

	Puntaje COMPAS									
Raza	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***
Afroamericana	365	346	298	337	323	318	343	301	317	227
Asiática	15	4	5	0	1	2	1	2	0	1
Caucásica	605	321	238	243	200	160	113	96	77	50
Hispana	159	89	73	47	39	27	28	14	17	16
Nativo americana	0	2	1	0	0	2	2	0	2	2
Otra	142	60	32	39	19	20	9	7	7	8

Tabla A.1.1. Distribución de cada puntaje COMPAS por razas.

	Puntaje COMPAS (reincidencia violenta)									
Raza	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***
Afroamericana	395	294	278	233	185	191	139	87	79	37
Asiática	15	2	1	4	1	2	0	1	0	0
Caucásica	657	221	193	126	94	72	46	20	24	6
Hispana	155	60	44	25	31	20	9	7	3	1
Nativo americana	2	1	0	2	0	1	0	0	1	0
Otra	116	44	27	18	14	14	9	7	3	3

Tabla A.1.2. Distribución de cada puntaje COMPAS por razas (reincidencia violenta).

VARIABLE BINARIA DE PUNTAJE COMPAS ALTO

Una de las críticas por parte de W. Flores et al. (2016) hacia el estudio de ProPublica (Larson et al., s/f), fue que al considerar los modelos de regresión logística para reincidencia y reincidencia violenta, se usó como variable de salida binaria una que distingue entre puntaje alto y bajo, comprimiendo los 10 niveles decimales del puntaje COMPAS de esta manera, y asignando todos los puntajes medios como puntajes altos. ProPublica justificó su decisión con una sección del estudio presentado por los mismos desarrolladores de COMPAS, Nothpointe, donde sugieren que puntajes medios y altos llaman más la atención de las agencias supervisoras. Aún así, W. Flores et al. (2016) sugieren que debería explorarse el resultado, al considerar los puntajes medios como bajos al convertirlos a la variable binaria, aunque no lo llevan a cabo en su estudio.

A continuación se muestran los resultados para el modelo de regresión logística planteado por ProPublica, pero en este caso considerando los puntajes medios como bajos, en vez de altos, como se considera en el trabajo original. También se exponen los resultados de la implementación original para facilitar la comparación.

Coeficientes de Regresión Logística (componentes del vector **w** y valor de *b*)		
Atributo correspondiente en el vector **x**	*Implementación Original (uniendo puntajes medios y altos)*	*Implementación Modificada (uniendo puntajes medios y bajos)*
Ninguno (el coeficiente es la constante *b*)	-1.52554	-3.02402
Factor de género: femenino (*gender_factor*)	0.22127	-0.12906
Factor de edad: mayor de 45 (*age_factor*)	-1.35563	-1.56431
Factor de edad: menor de 25 (*age_factor*)	1.30839	1.11849
Factor de raza: afroamericano (*race_factor*)	0.47721	0.55077
Factor de raza: asiático (*race_factor*)	-0.25441	0.33544
Factor de raza: hispano (*race_factor*)	-0.42839	-0.31666
Factor de raza: nativo americano (*race_factor*)	1.39421	1.23472
Factor de raza: otro (*race_factor*)	-0.82635	-0.66541
Conteo de antecedentes (*priors_count*)	0.26895	0.187363
Factor de crimen: delito menor (*crime_factor*)	-0.31124	-0.37893
Reincidencia en dos años (*two_year_recid*)	0.68586	0.83153

Tabla A.2.1.1. Comparación de coeficientes de regresión logística para reincidencia general, entre considerar los puntajes medios como bajos, y al considerarlos como altos.

	Implementación Original (uniendo puntajes medios y altos)				*Implementación Modificada (uniendo puntajes medios y bajos)*			
	RR	*RD*	*RC*	*OR*	*RR*	*RD*	*RC*	*OR*
Raza negra	1.452838	0.080898	0.901506	1.611567	1.677463	0.031402	0.967072	1.734580
Género Femenino	1.194795	0.0348	0.957631	1.247656	0.883885	-0.005382	1.005644	0.878925
Edad menor a 25 años	2.49612	0.267278	0.674588	3.700214	2.79345	0.083131	0.912828	3.060218

Tabla A.2.1.2. Comparación de medidas de la discriminación para reincidencia general, entre considerar los puntajes medios como bajos, y al considerarlos como altos.

Era de esperarse que los coeficientes de regresión logística y por ende las medidas de discriminación cambiaran con esta modificación, pues el problema queda estructurado de forma distinta: los elementos de puntaje medio, que el modelo antes percibía con un puntaje binario alto, ahora son percibidos con un puntaje binario bajo.

Sin embargo, los resultados en todo caso sustentan las conclusiones de ProPublica, pues al estructurar de esta forma el problema, los elementos de raza negra tienen una probabilidad aún mayor de tener puntaje alto, casi un 68% comparado con el 45% reportado en el trabajo original. Del mismo modo, los elementos de menos de 25 años ahora tienen una probabilidad casi 2.8 veces mayor, comparada a la probabilidad 2.5 veces mayor del trabajo original. El único resultado que cambió fue que los elementos de género femenino ahora tienen una probabilidad 11.6% menor de recibir un puntaje alto, mientras que en el trabajo original tenían una probabilidad 19.4% mayor, resultado que curiosamente los autores de ProPublica denominaron "sorpresivo" dado que en general las mujeres tienen menores niveles de criminalidad (Larson et al., s/f).

Ya que el riesgo relativo entre hombres y mujeres no es discutido más allá de presentar el resultado (ni siquiera se obtiene esta medida para el estudio sobre reincidencia violenta), se puede decir que las conclusiones de ProPublica no se verían afectadas por la modificación sugerida en (W. Flores et al., 2016). No se debe perder de vista que su estudio siempre estuvo enfocado en las diferencias entre elementos de razas afroamericana y caucásica, y sólo se presentan conclusiones con respecto a ello en el artículo editorial (Julia Angwin, 2016).

A continuación se presentan los resultados para reincidencia violenta, para observar también las diferencias en los resultados de este caso, aplicando la misma modificación.

Coeficientes de Regresión Logística (componentes del vector **w** y valor de *b*)		
Atributo correspondiente en el vector **x**	*Resultados de la reproducción (uniendo puntajes medios y altos)*	*Resultados de la reproducción (uniendo puntajes medios y bajos)*
Ninguno (el coeficiente es la constante *b*)	-2.24273	-3.12721
Factor de género: femenino (*gender_factor*)	-0.7289	-0.19867
Factor de edad: mayor de 45 (*age_factor*)	-1.74208	-1.50155
Factor de edad: menor de 25 (*age_factor*)	3.14591	1.24684
Factor de raza: afroamericano	0.65893	0.58611

(*race_factor*)		
Factor de raza: asiático (*race_factor*)	-0.98521	-0.0584
Factor de raza: hispano (*race_factor*)	-0.06416	-0.13387
Factor de raza: nativo americano (*race_factor*)	0.44793	1.47143
Factor de raza: otro (*race_factor*)	-0.20543	-0.49421
Conteo de antecedentes (*priors_count*)	0.13764	0.19389
Factor de crimen: delito menor (*crime_factor*)	-0.16367	-0.35454
Reincidencia en dos años (*two_year_recid*)	0.93448	1.14892

Tabla A.2.2.1. Comparación de coeficientes de regresión logística para reincidencia general, entre considerar los puntajes medios como bajos, y al considerarlos como altos (reincidencia violenta).

	Implementación Original (uniendo puntajes medios y altos)				*Implementación Modificada (uniendo puntajes medios y bajos)*			
	RR	*RD*	*RC*	*OR*	*RR*	*RD*	*RC*	*OR*
Raza negra	1.773927	0.074280	0.917834	1.932732	1.738782	0.031028	0.967612	1.796983
Género Femenino	0.507656	1.052271	-0.047254	0.482438	0.826073	-0.007305	1.007625	0.819822
Edad menor a 25 años	7.414217	0.615623	0.319018	23.240731	3.15120	0.090348	0.905692	3.479330

Tabla A.2.2.2. Comparación de medidas de la discriminación para reincidencia general, entre considerar los puntajes medios como bajos, y al considerarlos como altos (reincidencia violenta).

Nuevamente, para las medidas de discriminación, en particular el riesgo relativo (RR), se puede observar que la conclusión referente a las diferencias raciales no cambiaría, pues aunque originalmente se concluyó que un elemento de raza negra tiene una probabilidad 77.3% mayor de ser clasificado con un puntaje alto para reincidencia violenta, con la modificación esta probabilidad apenas baja a 73.8%, por lo que difícilmente cambiaría las conclusiones generales de ProPublica. El cambio para elementos menores de 25 años fue más drástico, y en vez de tener una probabilidad 7.4 veces mayor como en el trabajo original, tienen una probabilidad aproximadamente 3.2 veces mayor, pero este resultado tampoco tiene mayor relevancia para las conclusiones presentadas, pues ProPublica sólo menciona al respecto que la influencia de la edad al predecir el resultado, es aún más fuerte que la de la raza, lo cual sigue siendo verdadero. De la misma forma, el factor de género femenino representa una reducción de tan sólo 17.4% en la probabilidad de obtener un puntaje alto para reincidencia violenta, cuando en el modelo original la reducción era de casi un 50%, pero nuevamente, no tiene relevancia, pues ProPublica ni siquiera hace mención al factor de género en el estudio de la reincidencia violenta.

MODIFICACIONES DE LA VARIANZA A PRIORI

Al implementar el método bayesiano para el estudio de ProPublica, y las demás implementaciones, se mencionó que la varianza unitaria a priori fue adecuada, y no había necesidad de considerar una mayor. En estos casos, con los resultados obtenidos para los coeficientes, ilustrados en la figuras con las gráficas de sus distribuciones a posteriori, se nota que la varianza unitaria es suficientemente grande, pues asigna una probabilidad previa a los valores aprendidos de cada coeficiente para que tengan probabilidades a posteriori adecuadas al terminar el entrenamiento.

Se puede ilustrar que esto es cierto de varias maneras, pero aquí simplemente se repite el procedimiento asignando varianzas a priori mucho mayores, para los modelos de ProPublica de reincidencia general y violenta. A continuación se presentan las medias y varianzas a posteriori de estos modelos, para distintas varianzas a priori (hay que tener en cuenta que la media a priori sigue siendo 0 en todos los casos):

Medias y Varianzas a Posteriori de los Coeficientes de Regresión Logística para Diferentes Varianzas a Priori								
Dato correspondiente al vector **x**	**Resultados a posteriori para diferentes varianzas a priori**							
	varianza = 1 (resultado original)		*varianza = 10*		*varianza = 100*		*varianza = 1000*	
	Media	*Varianza*	*Media*	*Varianza*	*Media*	*Varianza*	*Media*	*Varianza*
Ninguno (el coeficiente es la constante *b*)	-1.489185	0.001074	-1.496395	0.001075	-1.497159	0.001075	-1.4973	0.001075
Factor de género: femenino (*gender_factor*)	0.270534	0.007617	0.27444	0.007664	0.274849	0.007669	0.274908	0.00767
Factor de edad: mayor de 45 (*age_factor*)	-1.415583	0.010041	-1.424497	0.010151	-1.425394	0.010163	-1.425475	0.010164
Factor de edad: menor de 25 (*age_factor*)	1.280887	0.002522	1.2891	0.002529	1.289936	0.002529	1.290024	0.002529
Factor de raza: afroamericano (*race_factor*)	0.448626	0.001319	0.449765	0.001322	0.449907	0.001322	0.449974	0.001323
Factor de raza: asiático (*race_factor*)	-0.346765	0.160633	-0.397031	0.185763	-0.401588	0.18855	-0.400408	0.18864
Factor de raza: hispano (*race_factor*)	-0.464798	0.015907	-0.470031	0.016096	-0.470537	0.016115	-0.470528	0.016117
Factor de raza: nativo americano (*race_factor*)	0.726923	0.460661	0.926894	0.657504	0.951121	0.685654	0.956091	0.689071
Factor de raza: otro (*race_factor*)	-0.799635	0.016925	-0.818166	0.01726	-0.820049	0.017294	-0.820176	0.017297
Conteo de antecedentes (*priors_count*)	0.254689	0.000136	0.255613	0.000137	0.255708	0.000137	0.255719	0.000137
Factor de crimen: delito menor *(crime_factor)*	-0.321789	0.002532	-0.32092	0.00254	-0.320838	0.002541	-0.320828	0.002541
Reincidencia en dos años (*two_year_recid*)	0.703154	0.001101	0.707115	0.001104	0.707527	0.001104	0.707583	0.001104

Tabla. A.3.1 Coeficientes obtenidos para la implementación bayesiana de ProPublica, con distintas varianzas a priori para el caso de reincidencia general.

Medias y Varianzas a Posteriori de los Coeficientes de Regresión Logística para Diferentes Varianzas a Priori								
Dato correspondiente al vector **x**	**Resultados a posteriori para diferentes varianzas a priori**							
	varianza = 1 (resultado original)		*varianza = 10*		*varianza = 100*		*varianza = 1000*	
	Media	*Varianza*	*Media*	*Varianza*	*Media*	*Varianza*	*Media*	*Varianza*
Ninguno (el coeficiente es la constante *b*)	-2.085440	0.003090	-2.108475	0.003119	-2.110889	0.003122	-2.111316	0.003122
Factor de género: femenino (*gender_factor*)	-0.704297	0.009678	-0.719471	0.009808	-0.721031	0.009821	-0.721114	0.009822
Factor de edad: mayor de 45 (*age_factor*)	-1.763843	0.025178	-1.808327	0.026533	-1.812993	0.026677	-1.81341	0.026692
Factor de edad: menor de 25 (*age_factor*)	3.046187	0.005902	3.09039	0.005975	3.094962	0.005983	3.095387	0.005984
Factor de raza: afroamericano (*race_factor*)	0.597401	0.006563	0.605478	0.006628	0.606356	0.006634	0.60663	0.006635
Factor de raza: asiático (*race_factor*)	-0.679483	0.335395	-0.911622	0.440253	-0.938235	0.453477	-0.935609	0.453536
Factor de raza: hispano (*race_factor*)	-0.247318	0.016879	-0.251213	0.017337	-0.251569	0.017384	-0.251392	0.017388
Factor de raza: nativo americano (*race_factor*)	-0.008983	0.535349	-0.061328	0.926125	-0.067181	0.991605	-0.060201	0.997476
Factor de raza: otro (*race_factor*)	-0.344551	0.092972	-0.359287	0.099057	-0.360797	0.099706	-0.360728	0.099768
Conteo de antecedentes (*priors_count*)	0.126147	0.000145	0.128183	0.000146	0.128395	0.000146	0.128416	0.000146
Factor de crimen: delito menor *(crime_factor)*	-0.131252	0.009287	-0.124864	0.009469	-0.124204	0.009488	-0.124137	0.009489
Reincidencia en dos años (*two_year_recid*)	0.926821	0.007675	0.941791	0.007813	0.94336	0.007827	0.943513	0.007826

Tabla. A.3.2 Coeficientes obtenidos para la implementación bayesiana de ProPublica, con distintas varianzas a priori, para el caso de reincidencia violenta.

Como se puede observar de los resultados, las varianzas y las medias a posteriori llegan a valores muy similares a los obtenidos originalmente usando varianza unitaria[91]. La cercanía entre las medias indica que no hay necesidad de incrementar la varianza, pues con distribuciones a priori más amplias se llegó técnicamente a los mismos valores, o por lo menos a valores muy cercanos que siguen estando dentro del rango cubierto por la distribución a priori de varianza unitaria. Del mismo modo, las varianzas a posteriori son casi iguales para todos los coeficientes, así que tener varianzas tan grandes en este caso no sería tampoco un problema, pues los resultados muestran que aún con una varianza mil veces mayor, tras observar los datos, la varianza a posteriori es apenas mayor que en el caso con varianza unitaria. Para esta prueba se consideraron sólo los modelos para el estudio de ProPublica, ya que las implementaciones del estudio de W. Flores et al. (2016) llevan a coeficientes en un rango similar e incluso más reducido; en particular, ningún coeficiente en esos modelos tiene una media a posteriori más alejada del cero que el coeficiente para el factor de edad menor a 25 años del modelo de reincidencia violenta del caso de ProPublica. Así que no hay razón

91 Nótese que los cambios más significativos son para coeficientes que tienen varianzas a posteriori elevadas para cualquier caso, es decir, los factores de razas asiática y nativo americana.

para pensar que la media sea suficiente para los coeficientes analizados aquí y no para los del otro estudio.

Del mismo modo podrían seleccionarse otras medias a priori, si hay razón para creer de antemano cuáles podrían ser los valores de cada coeficiente, pero como en este caso no se busca incorporar ningún conocimiento previo, esta posibilidad no se explorará en este trabajo[92].

DIFERENCIAS POR CAMBIOS EN LA OPTIMIZACIÓN

En las implementaciones de capítulos anteriores, se mencionó que las diferencias entre resultados de implementaciones frecuentistas y bayesianas no se debe exclusivamente a la elección de metodología, sino que también se debe a los cambios en el esquema de optimización, pues todos los modelos bayesianos se implementaron con el optimizador Adam, mientras que los frecuentistas usaron IWLS. Se mencionó que un método frecuentista puede usar el optimizador Adam, y entonces se pueden esperar valores diferentes para los valores puntuales del método frecuentista, que pueden ser más cercanos o lejanos de la media bayesiana que cuando se usan métodos de optimziación diferentes.

Para ilustrar esto, a continuación se presentará una comparación de coeficientes de los modelos implementados para el análisis de ProPublica, entre resultados obtenidos con el método bayesiano, con el método frecuentista usando IWLS, y con el método frecuentista usando Adam y las demás características de entrenamiento para modelos bayesianos (como el uso de mini-lotes). En el método frecuentista con Adam, la divergencia KL no forma parte del costo, así que se optimiza sólo la NLL, como al usar IWLS. Ya que en esta sección el único interés es comparar las diferencias en los coeficientes, para el método bayesiano sólo se reportan las medias a posteriori.

	Coeficientes de Regresión Logística (componentes del vector **w** y valor de *b*)		
Atributo correspondiente en el vector **x**	*Método Frecuentista (IWLS)*	*Método Frecuentista (Adam)*	*Método Bayesiano (Adam)*
Ninguno (el coeficiente es la constante *b*)	-1.52554	-1.49295	-1.489185
Factor de género: femenino (*gender_factor*)	0.22127	0.201568	0.270534
Factor de edad: mayor de 45 (*age_factor*)	-1.35563	-1.441086	-1.415583
Factor de edad: menor de 25 (*age_factor*)	1.30839	1.298927	1.280887
Factor de raza: afroamericano (*race_factor*)	0.47721	0.490852	0.448626
Factor de raza: asiático (*race_factor*)	-0.25441	-0.296864	-0.346765
Factor de raza: hispano (*race_factor*)	-0.42839	-0.416649	-0.464798
Factor de raza: nativo americano (*race_factor*)	1.39421	1.111446	0.726923

92 De cualquier modo, las variaciones en la media siguen siendo modificaciones de la distribución a priori, por lo que si las distribuciones siguen teniendo rangos congruentes con los datos observados, independientemente de estar centradas o no en cero, aún podrían encontrarse distribuciones a posteriori similares a aquellas que resultaron al usar una media igual a cero.

Factor de raza: otro (*race_factor*)	-0.82635	-0.815465	-0.799635
Conteo de antecedentes (*priors_count*)	0.26895	0.277748	0.254689
Factor de crimen: delito menor (*crime_factor*)	-0.31124	-0.272156	-0.321789
Reincidencia en dos años (*two_year_recid*)	0.68586	0.732739	0.703154

Tabla A.4.1. Comparaciones de coeficientes de regresión logística por método, para el modelo de reincidencia general de ProPublica.

	Coeficientes de Regresión Logística (componentes del vector **w** y valor de *b*)		
Atributo correspondiente en el vector **x**	*Método Frecuentista (IWLS)*	*Método Frecuentista (Adam)*	*Método Bayesiano (Adam)*
Ninguno (el coeficiente es la constante *b*)	-2.24273	-2.204032	-2.085440
Factor de género: femenino (*gender_factor*)	-0.7289	-0.703153	-0.704297
Factor de edad: mayor de 45 (*age_factor*)	-1.74208	-1.752526	-1.763843
Factor de edad: menor de 25 (*age_factor*)	3.14591	3.111013	3.046187
Factor de raza: afroamericano (*race_factor*)	0.65893	0.652026	0.597401
Factor de raza: asiático (*race_factor*)	-0.98521	-0.932884	-0.679483
Factor de raza: hispano (*race_factor*)	-0.06416	-0.159535	-0.247318
Factor de raza: nativo americano (*race_factor*)	0.44793	0.183418	-0.008983
Factor de raza: otro (*race_factor*)	-0.20543	-0.266942	-0.344551
Conteo de antecedentes (*priors_count*)	0.13764	0.140417	0.126147
Factor de crimen: delito menor (*crime_factor*)	-0.16367	-0.106616	-0.131252
Reincidencia en dos años (*two_year_recid*)	0.93448	0.937225	0.926821

Tabla A.4.2. Comparaciones de coeficientes de regresión logística por método, para el modelo de reincidencia violenta de ProPublica.

Los resultados comprueban que cambiar la optimización lleva a resultados diferentes, aún si se mantiene la misma función de costo, pues los resultados obtenidos por IWLS y usando el optimizador Adam, aún si los modelos son frecuentistas y usan la NLL como función de costo, pueden llegar a resultados diferentes. Como era de esperarse, usar el mismo optimizador Adam para caso frecuentista y bayesiano, no implica que se aprendan a medias a posteriori iguales a los valores

puntuales frecuentistas; esto tiene sentido porque las funciones de costo son distintas: sólo el método bayesiano minimiza la divergencia KL.

En los resultados también se puede apreciar que la relación entre los tres métodos no es tan sencilla: sería intuitivo pensar que un método frecuentista, con el mismo optimizador del método bayesiano, llegaría a valores que son un punto medio entre ambos resultados originales. Sin embargo, aunque esto sí ocurre en algunos casos, varía en qué tan cerca está el valor de cada uno de los otros dos, y en ocasiones incluso se encontraron valores más cercanos entre el método frecuentista usando IWLS y la media bayesiana que entre el método frecuentista con Adam y cualquiera de los dos (como en los coeficientes para el factor de edad mayor a 45 años y factor de crimen de delito menor).

El propósito de esta sección es simplemente mostrar que en este trabajo hay variaciones entre los resultados de cada metodología, que dependen sólo del método de optimización, y no de si la implementación es frecuentista o bayesiana, por lo que queda de más estudiar las potenciales diferencias para las implementaciones referentes al trabajo de W. Flores et al. (2016). Para obtener conclusiones más generales con respecto a la naturaleza de estas variaciones, sería necesario un análisis más cuidadoso de las diferencias en resultados y de los detalles en el entrenamiento de cada modelo.

RENDIMIENTO DE MODELOS CLASIFICADORES

Ya que este trabajo se enfoca en una comparación entre metodologías, específicamente en la complejidad y profundidad de los modelos usados, no se hizo énfasis en las típicas comparaciones de modelos clasificadores en términos de precisión y generalización usadas en el aprendizaje computacional.

En el análisis de ProPublica (*propublica/compas-analysis*, 2016/2020), la función de R que usaron para entrenar los modelos provee automáticamente los valores para el criterio AIC para los modelos de regresión logística, que como se mencionó en dicho capítulo, comparan el rendimiento de un modelo con el modelo nulo, que es un modelo sin atributos aparte de la constante *b* presente en todo modelo. El criterio AIC también considera el número de parámetros usados en el modelo, penalizando la complejidad del modelo. Este criterio se basa en la verosimilitud logarítmica (LL), y usa dos medidas considerando la LL del modelo nulo (LL_0), y la LL del modelo evaluado (LL_m): desviación nula, igual a $-2LL_0$, y desviación residual, igual a $-2LL_m$. El valor del criterio AIC es $-2(LL_m + k)$, donde k es el número de parámetros en el modelo (es decir, atributos, incluyendo la constante *b*, por lo que para el estudio de ProPublica se tiene $k = 12$) (*Akaike Information Criterion - an overview | ScienceDirect Topics*, s/f). Así como se obtuvo y se usó la LL para calcular las medidas presentes en el análisis de W. Flores et al. (2016), se puede hacer lo mismo con las medidas del AIC del estudio de ProPublica. Como se obtuvieron valores idénticos en los coeficientes para la reproducción frecuentista, era de esperarse que los valores del AIC también coincidieran:

	Reincidencia General		**Reincidencia Violenta**	
Medida	*ProPublica*	*Reproducción*	*ProPublica*	*Reproducción*
Desviación nula	8483.3	8483.4	4731.8	4731.8
Desviación Residual	6168.4	6168.4	2998.8	2998.8
AIC	6192.4	6192.4	3022.8	3022.8

Tabla A.5.1.1. Comparación de las medidas del criterio AIC entre el trabajo original de ProPublica y la reproducción.

Como se puede apreciar, las medidas son idénticas a las presentadas por ProPublica. Del mismo modo que se hizo con las medidas usadas por W. Flores et al. (2016), se pueden obtener los valores

del criterio AIC para las implementaciones bayesianas, y como se hizo en ese caso, nuevamente se usan únicamente las medias a posteriori de los modelos, pues como se mencionó en ese capítulo, son los valores que se usarían para las predicciones óptimas.

Medida	**Reincidencia General**	**Reincidencia Violenta**
Desviación nula	8484.3	4731.9
Desviación Residual	6173.2	3003.3
AIC	6197.2	3027.3

Tabla A.5.1.2. Medidas del criterio AIC para las implementaciones bayesianas (usando la media a posteriori).

Los valores para el método bayesiano son similares a los del método frecuentista. Estos valores no se calcularon ni se estudiaron en el cuerpo principal del texto, porque en el estudio de ProPublica nunca los usan para derivar conclusiones con respecto al sesgo. La única conclusión implícita que podría obtenerse es que el modelo postulado tiene un mejor ajuste a los datos que el modelo nulo, aún penalizando el número de parámetros, pero en cualquier caso no tiene relación con el sesgo racial.

Por su parte, las medidas del rendimiento predictivo de los cuatro modelos en el trabajo de W. Flores et al. (2016), que se implementan para reincidencias general y violenta, sí se usan para obtener algunas conclusiones relevantes, pero su propósito es relativo al método: sólo se usan para comparar los cuatro modelos de una misma implementación entre sí. Comparar el rendimiento entre metodología bayesiana y frecuentista requiere un análisis más cuidadoso que simplemente comparar las medidas numéricas, ya sea AIC, LL, chi cuadrada o pseudo-R^2 [93]. Además, estas medidas no dicen nada con respecto a la generalización de los modelos, pues evalúan el ajuste de cada modelo a los datos que usó, pero no proveen información acerca de cómo rendirán los modelos con nuevos datos que no se han visto o usado en el entrenamiento. Por eso, en esta sección se considerará otro tipo de medida.

Es necesario tener una medida de cuán acertado es un modelo aprendido, pues de lo contrario sus resultados no llevarán a conclusiones significativas. El modelo aprende una función $f(\mathbf{x})$ para relacionar las entradas $\mathbf{x}$ con la respuesta y. En la regresión logística, esta función es la función sigmoide, parametrizada con los coeficientes $\mathbf{w}$ y la constante b, cuyos valores son lo que se aprende durante el entrenamiento. A lo largo de este trabajo, estos coeficientes se usaron para derivar resultados y conclusiones en el contexto del sesgo racial en el aprendizaje computacional. Si la función aprendida no es adecuada y precisa para los datos que usa para entrenar, entonces sus resultados no podrían usarse para derivar este tipo de conclusiones, pues una función así no reflejaría la relación entre las entradas y la respuesta. Las medidas del AIC en (*propublica/compas-analysis*, 2016/2020), y las medidas de rendimiento en (W. Flores et al., 2016), son indicativas de que los modelos tienen un rendimiento adecuado en este sentido, en el contexto de comparaciones con modelos nulos, y en el caso de las implementaciones bayesianas y reproducciones, también se revisó que los modelos obtenidos fueran aceptables en este sentido.

Retomando el capítulo de clasificadores, ahí se expuso el término de generalización, que se refiere a que la función que aprenden los modelos con ejemplos fijos se generalice a nuevos ejemplos. En los casos de estudio presentados, y las implementaciones basadas en los mismos, el enfoque estaba en estudiar las funciones aprendidas para los datos, ya que las mismas (específicamente, los coeficientes de regresión logística) se interpretaron en términos de su significado en el contexto del sesgo racial. Sin embargo, los datos usados son el llamado *training set* o conjunto de entrenamiento, y como se menciona en el capítulo de clasificadores, la precisión

93 Por ejemplo, si se quiere comparar el valor para chi cuadrada de algún modelo con el de su contraparte bayesiana, también se debe tomar en cuenta que ambos usan modelos nulos distintos como referencia, el primero un modelo nulo frecuentista y el segundo uno bayesiano. Esto podría modificarse usando el modelo nulo frecuentista para ambos, por ejemplo, pero este tipo de decisiones deben planearse y justificarse para tener resultados significativos.

en este conjunto es irrelevante para la generalización, ya que el modelo está emitiendo predicciones con respecto a ejemplos que ya aprendió; además no es útil hacer estas predicciones, pues la respuesta para esas entradas ya se conoce (Murphy, 2012, p. 3).

Para considerar la generalización, la práctica común es usar un conjunto de entrenamiento, un conjunto de validación (*validation set*) y un conjunto de prueba (*test set*). El conjunto de entrenamiento se usa como se han usado todos los datos en este trabajo, para optimizar funciones de costo, reduciendo el **error de entrenamiento**, y llegar a valores para los parámetros (en este caso coeficientes de regresión logística) que tengan el mínimo error de entrenamiento[94]. El conjunto de prueba se usa para observar el rendimiento de un modelo en datos que no se usaron en el entrenamiento, y obtener el **error de generalización**[95], que describe la precisión de las predicciones para futuros datos. El conjunto de validación se usa para poder comparar modelos con base en su generalización, y elegir entre diversos valores de hiper-parámetros, sin cometer el error de elegir un modelo que tenga el mejor rendimiento por casualidad: si estas decisiones se tomaran con base en el conjunto de prueba, sería como ajustarlas para tener un buen rendimiento en el mismo, lo cual anula el propósito de separar los datos de entrenamiento de los datos de prueba (Murphy, 2012, pp. 22–23), (*MLPR w2a - Machine Learning and Pattern Recognition*, s/f).

En los modelos implementados en este trabajo, todos los datos disponibles para cada modelo se usan como un conjunto de entrenamiento. Para el estudio de ProPublica el punto es derivar conclusiones para los datos que recolectaron, interpretando los coeficientes de los modelos. Si no se considera la capacidad de generalización del modelo, entonces las conclusiones basadas en el mismo se limitan a los datos en cuestión, pues no hay ningún indicador de la capacidad de modelo para mostrar una relación precisa para datos recolectados de otros lugares o periodos. En el estudio de W. Flores et al. (2016) se puede decir lo mismo de las conclusiones basadas en el radio de posibilidades (OR), basado a su vez en los coeficientes encontrados para los modelos. Pero además, ellos obtienen conclusiones basadas en las diferencias de la capacidad predictiva, o rendimiento, de los modelos que implementan, y al obtener dicho rendimiento sin conjuntos de prueba o validación, su medida sólo indica qué tan bien se ajustan los modelos a los datos usados, sin contemplar su rendimiento para datos futuros.

El objetivo en esta sección no es incorporar el concepto de generalización a las conclusiones y análisis desarrollados, aunque sería una dirección interesante para estudiar el caso de estudio con otra perspectiva. El punto es simplemente hacer una comparación de las implementaciones bayesianas y frecuentistas, pero desde el punto de vista de estas métricas tan importantes y comunes en el aprendizaje computacional, diferente al enfoque conceptual que se desarrolla en el escrito.

Por lo general se busca usar cuantos datos sea posible para el entrenamiento. Es común dividir los datos en un 80% para entrenamiento y 20% para validación, considerando los datos de prueba aparte, u otra división posible es usar 80% para entrenamiento, 10% para validación, y 10% para prueba[96], cuando los datos de prueba se consideran también en el 100% de datos disponibles. Cuando los datos son escasos, otra opción es usar validación cruzada (*cross validation* o CV), donde los datos se dividen en K bloques, y el proceso de entrenamiento se repite K veces, cada vez usando un bloque distinto como conjunto de prueba y el resto como conjunto de entrenamiento, y el error de generalización se da como el promedio de las K repeticiones. Este proceso no se usará en este caso, pero cabe mencionar que no es sencillo derivar conclusiones estadísticamente rigurosas del rendimiento al usar CV, y si se pretende implementar, tanto un proceso de prueba, como uno de

94 El mínimo que se haya encontrado con los algoritmos implementados y sus características particulares.

95 En realidad, el error de prueba promedio, que es la cantidad calculada en este punto, es una aproximación del verdadero error de generalización, que es un concepto que involucra la probabilidad verdadera y desconocida de las respuestas (*MLPR w2a - Machine Learning and Pattern Recognition*, s/f).

96 Una limitación de usar datos de prueba que son sólo una división del conjunto total disponible, de donde también viene el conjunto de entrenamiento, es que a fin de cuentas los datos vienen de la misma distribución, y los datos que realmente sean del futuro tendrán otra distribución. Sin embargo el método es útil para hacer comparaciones cuantitativas y simples entre modelos (*MLPR w2a - Machine Learning and Pattern Recognition*, s/f).

validación, se tendría que usar un CV anidado, resultando en K^2 repeticiones para el entrenamiento, lo cual es problemático (Murphy, 2012, pp. 23–24), (*MLPR w2a - Machine Learning and Pattern Recognition*, s/f.)

En este caso simplemente se dividirán los datos en un 80% de entrenamiento, 10% de validación, y 10% de prueba. En realidad la validación no tendrá el rol que suele tener para tomar decisiones con respecto a los modelos, y más bien se usará para ilustrar el proceso de elegir un modelo con base en la validación y después revisar la elección con el conjunto de prueba. En este caso, ya no se presentarán los coeficientes de cada modelo, o los detalles de las distribuciones a posteriori aprendidas, y simplemente se obtendrán los errores durante el entrenamiento, la validación y la prueba de los modelos. Todos los modelos serán entrenados con los mismos métodos que en el trabajo principal, pero esta vez usando sólo el 80% de los datos que se designe para el entrenamiento. El error reportado será la verosimilitud logarítmica negativa promedio (NLL/D)[97], y también se calculará la precisión de los modelos, que es la fracción de las predicciones que resulten correctas. Como en las medidas usadas en el análisis de W. Flores et al. (2016), y en las relacionadas con el AIC en esta sección, para las implementaciones bayesianas se usan las medias a posteriori para todas las tareas de predicción. También se incluyen los resultados para los modelos frecuentistas entrenados mediante el optimizador Adam, para observar el efecto en estos resultados como se revisó en la sección del presente anexo referente a diferencias por cambios en la optimización.

Primero se presenta el rendimiento para los modelos de reincidencia general y violenta implementados por ProPublica.

	Reincidencia General		**Reincidencia Violenta**	
	Error (NLL/D)	*Precisión*	*Error (NLL/D)*	*Precisión*
Método Frecuentista (IWLS)	0.4937	0.757	0.3672	0.848
Método Frecuentista (Adam)	0.4947	0.755	0.3674	0.849
Método Bayesiano (Adam)	0.4946	0.760	0.3675	0.847

Tabla. A.6.1.1. Error de entrenamiento y precisión en el conjunto de entrenamiento para los modelos de ProPublica.

Para los datos de entrenamiento, el método frecuentista mediante IWLS logra el menor error, en reincidencia general y violenta, pero el método bayesiano tuvo la mejor precisión en reincidencia general y el método frecuentista con Adam tuvo la mejor precisión en reincidencia violenta. De cualquier modo, las diferencias en errores y precisiones son muy reducidas entre los tres métodos, y se puede decir que los tres encontraron relaciones equivalentemente efectivas para predecir las respuestas con los datos de entrenamiento.

Antes de pasar a la validación es importante mencionar que el error, o bien la NLL, es una métrica que considera las predicciones con mayor detalle que la precisión, pues para su cálculo se usan los valores de las probabilidades de Bernoulli que el modelo determina para cada entrada. Para calcular la precisión, estas probabilidades deben separarse con algún umbral para determinar a partir de qué probabilidad de éxito una salida es positiva o negativa, pues las predicciones son binarias y no continuas como las probabilidades. Por ejemplo, si para cierta entrada el modelo predice una

97 Donde D es la cantidad de datos en el conjunto en cuestión. Es más conveniente reportar errores promedio, pues tienen una interpretación más sencilla, y ya que el error, o en este caso la NLL, considera una suma de errores para cada registro, obtener el promedio permite comparar errores entre conjuntos de datos con distintas cantidades de registros (como suele ser entre conjunto de entrenamiento y conjuntos de prueba o validación) (*MLPR w2a - Machine Learning and Pattern Recognition*, s/f).

probabilidad de éxito de 0.227, la NLL considera qué tan buena fue la predicción exacta para la salida verdadera. Para la precisión es necesario establecer un umbral, para entonces predecir en términos binarios: para elementos con probabilidades debajo del umbral, se predice que la respuesta será negativa (un fracaso, o un cero lógico) y para las probabilidades por encima del mismo, lo contrario. Una opción simple para el umbral es 0.5, pues las probabilidades van de 0 a 1, y este es el umbral empleado para las precisiones calculadas en esta sección (en el ejemplo de la probabilidad de 0.227, con este umbral, se predice un resultado negativo o fracaso, pues está debajo de 0.5). Normalmente el umbral debe elegirse con más cuidado, porque depende de la aplicación.

Cabe mencionar que para todos los modelos de reincidencia violenta, ya sea que tenga como respuesta el puntaje COMPAS binario como en el primer caso de estudio, o la reincidencia en dos años (también binaria), como en el segundo caso, las respuestas positivas son escasas. Se confirmó que en todos grupos de datos de cada caso, las respuestas en casos de reincidencia violenta son negativas en más del 80% de los casos: esto quiere decir que un modelo que predice que no habrá reincidencia por parte de ningún elemento, o que el puntaje será bajo para todos los elementos, logrará una precisión de 0.8 o mayor. En estos casos, como se sabe de antemano que el evento a predecir es raro, se puede seleccionar un umbral diferente para las probabilidades que dé el modelo (lo mismo debería considerarse para eventos que son muy comunes).

En este trabajo se reporta la precisión porque es una medida común, y refleja el rendimiento cuando el modelo se ve forzado a emitir una decisión binaria, pero es importante notar sus limitaciones, en particular para este trabajo, donde se toma simplemente el valor de 0.5. El error tomará precedencia en las comparaciones entre modelos.

Ahora bien, ya que los modelos fueron entrenados y mostraron rendimientos aceptables, se pasa a la validación. En este caso esta fase es sólo ilustrativa, pues lo usual es entrenar los modelos comparando el rendimiento de distintos hiper-parámetros. Para el caso frecuentista se podría considerar que esto sí se hizo, pues el modelo se está entrenando con IWLS y con Adam, pero en el caso bayesiano también se podría hacer algo similar, como probar diferentes medias o varianzas a priori, o hacer alguna otra modificación en el proceso de aprendizaje.

	Reincidencia General		**Reincidencia Violenta**	
	Error (NLL/D)	*Precisión*	*Error (NLL/D)*	*Precisión*
Método Frecuentista (IWLS)	0.5257	0.741	0.3879	0.818
Método Frecuentista (Adam)	0.5260	0.741	0.3863	0.821
Método Bayesiano (Adam)	0.5278	0.739	0.3897	0.816

Tabla. A.6.1.2. Error de validación y precisión en el conjunto de validación para los modelos de ProPublica.

Viendo los errores de validación, el modelo frecuentista con IWLS tuvo el mejor rendimiento en reincidencia general, y el modelo frecuentista con Adam en reincidencia violenta. Con base en esto podría elegirse el modelo con IWLS para reincidencia general, y el modelo con Adam para reincidencia violenta. En el caso del modelo bayesiano no se hacen variaciones en los parámetros que entren en la validación, pero si se comparara el uso de diferentes varianzas a priori o cambios en el aprendizaje (como el optimziador o el uso de mini-lotes) podrían elegirse características específicas con base en la validación. Incluso el tipo de metodología podría ser evaluada mediante la validación, en cuyo caso, al tener aquí un peor rendimiento para el modelo bayesiano, que para los dos modelos frecuentistas, la elección sería la metodología frecuentista.

Finalmente los modelos seleccionados se evalúan con el error de generalización. En este caso se presentan las medidas para todos los modelos, pero en realidad en este punto sólo debería

evaluarse el modelo frecuentista con IWLS para reincidencia general y el modelo frecuentista con Adam para reincidencia violenta, ya que fueron los elegidos durante la validación.

	Reincidencia General		**Reincidencia Violenta**	
	Error (NLL/D)	*Precisión*	*Error (NLL/D)*	*Precisión*
Método Frecuentista (IWLS)	**0.5256**	**0.739**	0.4084	0.815
Método Frecuentista (Adam)	0.5279	0.739	**0.4099**	**0.820**
Método Bayesiano (Adam)	0.5240	0.747	0.4098	0.815

Tabla. A.6.1.3. Error de generalización (prueba) y precisión para los modelos de ProPublica.

En la fase de prueba el objetivo es cuantificar qué tan bien generalizarán los modelos seleccionados. En este caso no era de esperarse una gran diferencia con el error de validación, ya que en realidad no se probaron variantes significativas de los hiper-parámetros. Los modelos seleccionados se resaltan en negritas: para el modelo de reincidencia general, la generalización es bastante buena, tiene un error aún mejor que el de validación y que no es mucho mayor al de entrenamiento. El modelo de reincidencia violenta tampoco tiene una mala generalización, aunque su error aumenta más comparando al entrenamiento y la validación: mayores aumentos en el error de validación o prueba, en comparación con el de entrenamiento, indican un mayor grado de sobre ajuste (*over-fitting*)[98]. Si un modelo aprende una relación general suficientemente buena entre entradas y salidas durante el entrenamiento, los errores deberían ser similares[99].

Viendo los otros resultados, se puede apreciar que de haber seleccionado el modelo bayesiano para reincidencia general, y el frecuentista con IWLS para reincidencia violenta, se hubiera tenido la mejor generalización de este ejemplo. El punto es usar el error de generalización para evaluar los modelos seleccionados, no para escoger entre modelos, pues como ya se mencionó antes, esto puede llevar a elegir modelos que rinden bien en los datos de prueba por pura casualidad, y en general al ya mencionado *over-fitting* o sobre ajuste del modelo, que dará como resultado una mala generalización en datos futuros.

Presentar los resultados en este caso fue sencillo porque sólo hay dos modelos. Para el caso del estudio de W. Flores et al. (2016), se implementaron doce modelos, así que los resultados se separan para los modelos de reincidencia general, los de reincidencia violenta con datos filtrados, y los de reincidencia violenta con datos sin filtrar. Siguiendo el procedimiento para los modelos del caso de ProPublica, se seleccionará un modelo para cada caso, así que esta vez se seleccionará un total de doce modelos.

98 Aunque el aumento del error de validación puede evidenciar un mal ajuste, es importante notar que el error de validación y/o prueba es la medida de generalización, no su diferencia con el error de entrenamiento. Un modelo con mejor error de validación que otro indica mejor generalización, independientemente de si empeoró más con respecto a su propio error de entrenamiento (*MLPR w2a - Machine Learning and Pattern Recognition*, s/f).

99 Por lo general no se tendrán suficientes datos para tener un buen ajuste en todo el espacio de las entradas, así que tener forzosamente errores similares de entrenamiento y validación o prueba es un objetivo irrazonable, porque lógicamente el modelo será más preciso en los puntos cercanos a los datos de entrenamiento. De cualquier modo, analizar los errores provee información valiosa (*MLPR w2a - Machine Learning and Pattern Recognition*, s/f).

Error de Entrenamiento:

	Modelo							
	1		**2**		**3**		**4**	
	Error (NLL/D)	*Precisión*	*Error (NLL/D)*	*Precisión*	*Error (NLL/D)*	*Precisión*	*Error (NLL/D)*	*Precisión*
Método Frecuentista (IWLS)	0.6630	0.598	0.6144	0.666	0.6140	0.668	0.6140	0.668
Método Frecuentista (Adam)	0.6631	0.603	0.6153	0.665	0.6150	0.667	0.6164	0.664
Método Bayesiano (Adam)	0.6651	0.587	0.6153	0.666	0.6150	0.666	0.6151	0.67

Tabla. A.7.1.1. Error de entrenamiento y precisión en el conjunto de entrenamiento para los modelos de Flores et al. (2016) (reincidencia general).

	Modelo							
	1		**2**		**3**		**4**	
	Error (NLL/D)	*Precisión*	*Error (NLL/D)*	*Precisión*	*Error (NLL/D)*	*Precisión*	*Error (NLL/D)*	*Precisión*
Método Frecuentista (IWLS)	0.4222	0.833	0.3911	0.838	0.3902	0.837	0.3898	0.838
Método Frecuentista (Adam)	0.4471	0.833	0.4192	0.828	0.4185	0.823	0.4187	0.822
Método Bayesiano (Adam)	0.4225	0.833	0.3933	0.838	0.3930	0.838	0.3942	0.839

Tabla. A.7.1.2. Error de entrenamiento y precisión en el conjunto de entrenamiento para los modelos de Flores et al. (2016) (reincidencia violenta, datos filtrados)

	Modelo							
	1		**2**		**3**		**4**	
	Error (NLL/D)	*Precisión*	*Error (NLL/D)*	*Precisión*	*Error (NLL/D)*	*Precisión*	*Error (NLL/D)*	*Precisión*
Método Frecuentista (IWLS)	0.4349	0.826	0.4065	0.833	0.4060	0.834	0.4058	0.833
Método Frecuentista (Adam)	0.4349	0.826	0.4070	0.834	0.4068	0.833	0.4071	0.833
Método Bayesiano (Adam)	0.4390	0.826	0.4114	0.831	0.4108	0.833	0.4092	0.831

Tabla. A.7.1.3. Error de entrenamiento y precisión en el conjunto de entrenamiento para los modelos de Flores et al. (2016) (reincidencia violenta, datos sin filtrar).

Como en el caso anterior, los errores de entrenamiento reflejan que los tres métodos logran errores similares en los doce modelos, aunque el método frecuentista con IWLS domina en rendimiento a los otros dos, aún si las diferencias son pequeñas. No siempre consigue la mejor precisión, pero por lo mencionado antes, el error es más relevante para las comparaciones.

Error de validación:

	Modelo							
	1		**2**		**3**		**4**	
	Error (NLL/D)	*Precisión*	*Error (NLL/D)*	*Precisión*	*Error (NLL/D)*	*Precisión*	*Error (NLL/D)*	*Precisión*
Método Frecuentista (IWLS)	0.6690	0.604	0.6229	0.646	0.6256	0.648	0.6256	0.648
Método Frecuentista (Adam)	0.6698	0.608	0.6237	0.655	0.6258	0.653	0.6252	0.644
Método Bayesiano (Adam)	0.6717	0.583	0.6234	0.638	0.6261	0.648	0.6286	0.640

Tabla. A.7.2.1. Error de validación y precisión en el conjunto de validación para los modelos de Flores et al. (2016) (reincidencia general).

	Modelo							
	1		**2**		**3**		**4**	
	Error (NLL/D)	*Precisión*	*Error (NLL/D)*	*Precisión*	*Error (NLL/D)*	*Precisión*	*Error (NLL/D)*	*Precisión*
Método Frecuentista (IWLS)	0.4861	0.808	0.4538	0.802	0.4548	0.799	0.4559	0.808
Método Frecuentista (Adam)	0.4943	0.808	0.4742	0.796	0.4753	0.790	0.4820	0.775
Método Bayesiano (Adam)	0.4871	0.808	0.4598	0.802	0.4619	0.799	0.4675	0.799

Tabla. A.7.2.2. Error de validación y precisión en el conjunto de validación para los modelos de Flores et al. (2016) (reincidencia violenta, datos filtrados).

	Modelo							
	1		**2**		**3**		**4**	
	Error (NLL/D)	*Precisión*	*Error (NLL/D)*	*Precisión*	*Error (NLL/D)*	*Precisión*	*Error (NLL/D)*	*Precisión*
Método Frecuentista (IWLS)	0.4491	0.834	0.4046	0.834	0.4073	0.831	0.4075	0.829
Método Frecuentista (Adam)	0.4492	0.834	0.4069	0.831	0.4093	0.834	0.4088	0.829
Método Bayesiano (Adam)	0.4545	0.834	0.4135	0.831	0.4153	0.831	0.4133	0.826

Tabla. A.7.2.3. Error de validación y precisión en el conjunto de validación para los modelos de Flores et al. (2016) (reincidencia violenta, datos sin filtrar).

Usando los resultados de validación para seleccionar modelos como en el caso anterior, para casi todos los modelos el método frecuentista con IWLS tiene el menor error, así que es el elegido para once de los doce modelos, la excepción es el modelo 4 de reincidencia general, donde el método frecuentista con Adam tuvo el mejor rendimiento.

Error de generalización:

	Modelo							
	1		**2**		**3**		**4**	
	Error (NLL/D)	*Precisión*	*Error (NLL/D)*	*Precisión*	*Error (NLL/D)*	*Precisión*	*Error (NLL/D)*	*Precisión*
Método Frecuentista (IWLS)	**0.6619**	**0.632**	**0.6151**	**0.670**	**0.6152**	**0.672**	0.6152	0.672
Método Frecuentista (Adam)	0.6620	0.628	0.6159	0.674	0.6161	0.670	**0.6173**	**0.668**
Método Bayesiano (Adam)	0.6646	0.613	0.6155	0.666	0.6159	0.670	0.6172	0.664

Tabla. A.7.3.1. Error de generalización y precisión en el conjunto de prueba para los modelos de Flores et al. (2016) (reincidencia general).

	Modelo							
	1		**2**		**3**		**4**	
	Error (NLL/D)	*Precisión*	*Error (NLL/D)*	*Precisión*	*Error (NLL/D)*	*Precisión*	*Error (NLL/D)*	*Precisión*
Método Frecuentista (IWLS)	**0.4784**	**0.819**	**0.4115**	**0.834**	**0.4164**	**0.825**	**0.4189**	**0.816**
Método Frecuentista (Adam)	0.5015	0.819	0.4364	0.801	0.4422	0.795	0.4497	0.783
Método Bayesiano (Adam)	0.4779	0.819	0.4178	0.837	0.4219	0.828	0.4304	0.825

Tabla. A.7.3.2. Error de generalización y precisión en el conjunto de prueba para los modelos de Flores et al. (2016) (reincidencia violenta, datos filtrados).

	Modelo							
	1		**2**		**3**		**4**	
	Error (NLL/D)	*Precisión*	*Error (NLL/D)*	*Precisión*	*Error (NLL/D)*	*Precisión*	*Error (NLL/D)*	*Precisión*
Método Frecuentista (IWLS)	**0.4232**	**0.841**	**0.3913**	**0.848**	**0.3915**	**0.846**	**0.3917**	**0.848**
Método Frecuentista (Adam)	0.4230	0.841	0.3932	0.848	0.3937	0.848	0.3946	0.848
Método Bayesiano (Adam)	0.4296	0.841	0.4004	0.841	0.4003	0.843	0.3992	0.838

Tabla. A.7.3.3. Error de generalización y precisión en el conjunto de prueba para los modelos de Flores et al. (2016) (reincidencia violenta, datos sin filtrar).

Al evaluar la generalización de los modelos seleccionados (resaltados en negritas) en el conjunto de pruebas, para los cuatro modelos de reincidencia general se tienen resultados adecuados, pues los errores tienen cambios mínimos con respecto a los de entrenamiento, y en el caso del modelo 1, el rendimiento fue aún superior que durante el entrenamiento.

Para el caso de reincidencia violenta, con datos filtrados, hay incrementos más notables en el error, especialmente en el modelo 1, seguido del modelo 4. Aún así, los errores se reducen con respecto a la validación. Para la reincidencia violenta, con datos sin filtrar, en todos los modelos elegidos la generalización es bastante buena, logrando errores aún menores que los de entrenamiento.

Los procedimientos desarrollados en esta sección ilustran el uso de estas medidas para seleccionar modelos y evaluar su generalización. Estos procedimientos se llevaron acabo en un nivel muy superficial, especialmente durante la validación. La conclusión relevante es que los errores de los tres métodos son similares en las tres etapas de cada modelo, y por lo tanto, los análisis desarrollados en el cuerpo principal de la tesis no están sujetos a que algún método esté describiendo una relación más precisa o que generalice mejor para datos futuros.

Del mismo modo, se ha hablado de los compromisos o *trade-offs* entre metodologías bayesiana y frecuentista, recalcando que en este trabajo se lidia con ellos a un nivel conceptual. En las métricas obtenidas en esta sección se puede hablar de compromisos en términos del rendimiento:

por ejemplo, se podría añadir que además de la mayor complejidad en el diseño de los modelos bayesianos, también se hace un sacrificio en la precisión y generalización de los modelos obtenidos, pues en estos resultados los modelos con los menores errores siempre fueron frecuentistas[100].

Todo esto es sólo ilustrativo, porque en realidad las diferencias en estas medidas son insuficientes para hablar de un cambio significativo de rendimiento entre modelos. En general, la poca relevancia de las diferencias en esta sección, y en la mayoría de las conclusiones derivadas de los modelos por ambas metodologías, reflejan lo expresado en Gelman et al. (2013, p. 6): en muchos análisis simples, se derivan conclusiones superficialmente similares de ambos métodos.

Como se mencionó antes, algunos de los trabajos relacionados se enfocan en incorporar nociones de justicia en las funciones de costo a optimizar (Dimitrakakis et al., 2017), (Perrone et al., 2020). Si esto se desarrollara en este escrito, este tipo de comparación de rendimiento sería más relevante, pues serviría para estudiar cómo cambia el rendimiento de cada método al incorporar definiciones matemáticas de justicia en la optimización. Sin embargo, este tipo de desarrollo es propio de tareas de prevención de discriminación, y no tanto de descubrimiento, que es el caso de este escrito.

Tiempos de ejecución

Otra métrica común de comparación es el tiempo de ejecución. Aunque por lo general no está dentro del enfoque de trabajos de justicia en el aprendizaje computacional, puede visitarse en esta sección para ilustrar hasta qué grado el trabajo adicional para el método bayesiano mencionado en este trabajo, a nivel de diseño y herramientas empleadas, se traduce en diferencias en esta sencilla métrica[101].

Para estos resultados, deben recordarse las especificaciones del equipo usado para todas las implementaciones, ya mencionadas con anterioridad: procesador Intel Core-i7 (cuarta generación) 4810MQ de 2.8GHz, memoria RAM de 16GB con tecnología DDR3L SDRAM, y sistema operativo Windows 8.1. Los tiempos de ejecución presentados son estrictamente para el entrenamiento de los modelos, aislando el proceso de toda fase de carga o procesamiento de datos y librerías. Se presentan los tiempos de ejecución para los catorce modelos (los dos de ProPublica y los doce de W. Flores et al. (2016)). Se presentan los tiempos para el método bayesiano y la reproducción frecuentista, añadiendo también el caso frecuentista usando Adam en vez de IWLS para una comparación más completa. Todos los tiempos se expresan en milisegundos (*ms*).

	Modelo	
	Reincidencia General	*Reincidencia Violenta*
	Tiempo de ejecución (ms)	*Tiempo de ejecución (ms)*
Método Frecuentista (IWLS)	36.979	15.600
Método Frecuentista (Adam)	560.068	569.908
Método Bayesiano (Adam)	753.909	747.768

Tabla. A.8.1. Tiempos de ejecución para el entrenamiento de los modelos del análisis de ProPublica, por método.

100 Con estos resultados podría explorarse implementar ambos métodos, como se propone en el trabajo de McNair (2018), y no necesariamente elegir un sólo método durante la validación como se desarrolló en esta sección.

101 También debe recordarse que, como se mencionó al describir las herramientas empleadas en esta tesis en el capítulo 6, el tiempo de ejecución no es de particular relevancia para la presente aplicación, que en general involucra implementaciones relativamente simples.

	Modelo			
	1	**2**	**3**	**4**
	Tiempo de ejecución (ms)	*Tiempo de ejecución (ms)*	*Tiempo de ejecución (ms)*	*Tiempo de ejecución (ms)*
Método Frecuentista (IWLS)	9.866	9.597	9.480	9.660
Método Frecuentista (Adam)	581.224	568.725	559.223	574.78
Método Bayesiano (Adam)	709.122	707.208	720.223	721.449

Tabla. A.8.2.1. Tiempos de ejecución para el entrenamiento de los modelos del análisis de Flores et al. (2016), por método (reincidencia general).

	Modelo			
	1	**2**	**3**	**4**
	Tiempo de ejecución (ms)	*Tiempo de ejecución (ms)*	*Tiempo de ejecución (ms)*	*Tiempo de ejecución (ms)*
Método Frecuentista (IWLS)	13.475	13.258	12.497	12.797
Método Frecuentista (Adam)	568.127	528.007	562.313	569.035
Método Bayesiano (Adam)	726.842	718.37	728.671	731.092

Tabla. A.8.2.2. Tiempos de ejecución para el entrenamiento de los modelos del análisis de Flores et al. 2016, por método (reincidencia violenta, datos filtrados).

	Modelo			
	1	**2**	**3**	**4**
	Tiempo de ejecución (ms)	*Tiempo de ejecución (ms)*	*Tiempo de ejecución (ms)*	*Tiempo de ejecución (ms)*
Método Frecuentista (IWLS)	12.676	13.196	13.634	15.472
Método Frecuentista (Adam)	557.833	564.035	566.029	565.153
Método Bayesiano (Adam)	709.600	710.923	711.428	724.642

Tabla. A.8.2.3. Tiempos de ejecución para el entrenamiento de los modelos del análisis de Flores et al., por método (reincidencia violenta, datos sin filtrar).

En los tiempos de ejecución, efectivamente se refleja lo que se sugirió en las etapas de diseño de modelos en ambos análisis desarrollados y en el escrito en general: la implementación bayesiana es más compleja en este caso, y no sólo en su estructura conceptual, sino que el ajuste de los modelos requiere operaciones computacionalmente más costosas. Sin embargo, como se puede ver en los

tiempos de ejecución para el método frecuentista con Adam, gran parte del aumento en tiempo de ejecución es atribuible al esquema de optimización, pues aunque el método sea frecuentista, usar el mismo optimizador que se usa para el método bayesiano implica un aumento temporal considerable. Sin embargo, aún hay una marcada diferencia entre métodos aunque usen la misma optimización, pues el método bayesiano incorpora la divergencia KL en su costo, y evidentemente resulta en ejecuciones más largas[102].

En conclusión, el análisis del rendimiento en términos de errores y generalización, muestra que no hay una diferencia relevante entre modelos en cuanto a su ajuste a los datos y su rendimiento con datos nuevos. El análisis en términos de tiempos de ejecución sustenta lo que se planteó a través del escrito: el método bayesiano puede ofrecer respuestas más completas e intuitivas con su cuantificación de la incertidumbre, pero a su vez implica modelos más elaborados, no sólo en concepto y diseño, también en implementación y ejecución.

REFERENCIAS

Akaike Information Criterion—An overview | ScienceDirect Topics. (s/f). Recuperado el 28 de junio de 2020, de https://www.sciencedirect.com/topics/medicine-and-dentistry/akaike-information-criterion

Dimitrakakis, C., Liu, Y., Parkes, D., & Radanovic, G. (2017). *Bayesian fairness*.

Gelman, A., Stern, H. S., Carlin, J. B., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. Chapman and Hall/CRC.

Julia Angwin, J. L. (2016, mayo 23). *Machine Bias* [Text/html]. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (s/f). *How We Analyzed the COMPAS Recidivism Algorithm*. ProPublica. Recuperado el 2 de mayo de 2020, de https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

McNair, D. S. (2018). Preventing Disparities: Bayesian and Frequentist Methods for Assessing Fairness in Machine-Learning Decision-Support Models. En M. S. F. Nezhad (Ed.), *New Insights into Bayesian Inference*. IntechOpen. https://doi.org/10.5772/intechopen.73176

MLPR w2a—Machine Learning and Pattern Recognition. (s/f). Recuperado el 30 de junio de 2020, de https://www.inf.ed.ac.uk/teaching/courses/mlpr/2018/notes/w2a_train_test_val.html

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT press.

102 Hay que tener en cuenta que a pesar de tener resultados que pueden reproducirse para el resto del trabajo, los tiempos de ejecución son una medida más inestable y dependiente del equipo y sus procesos internos. En este caso, fue fácil derivar conclusiones por las claras tendencias en cada método entre los distintos modelos, pero para generalizar las conclusiones debidamente, o estudiar cualidades más específicas con respecto al tiempo de ejecución, se requeriría mucho más rigor experimental.

Perrone, V., Donini, M., Kenthapadi, K., & Archambeau, C. (2020). *Fair Bayesian Optimization*.

Propublica/compas-analysis. (2020). [Jupyter Notebook]. ProPublica. https://github.com/propublica/compas-analysis (Original work published 2016)

W. Flores, A., Bechtel, K., & Lowenkamp, C. (2016). False Positives, False Negatives, and False Analyses: A Rejoinder to “Machine Bias: There’s Software Used Across the Country to Predict Future Criminals. And it’s Biased Against Blacks.” *Federal probation, 80*.