Universidad Nacional Autónoma de México

Facultad de Ciencias

algunos métodos de aprendizaje supervisado

T E S I S

P R E S E N T A :

Gerardo Martínez Mejía

QUE PARA OBTENER EL TÍTULO DE:

Matemático

TUTORA

Dra. Guillermina Eslava Gómez

Ciudad Universitaria, CD. MX., 2023

Universidad Nacional Autónoma de México
UNAM
Biblioteca Central
Dirección General de Bibliotecas de la UNAM

UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

Para mi mamá, mi papá, mi hermano
y los que me acompañaron en este viaje

Agradecimientos

Aprovecho este espacio para expresar mi gratitud y aprecio a todas las personas que me acompañaron en esta etapa de mi vida.

Primeramente agradezco a mis padres y mi hermano por todo el apoyo y cariño incondicional que me han dado. Les agradezco por el apoyo en mis tomas de decisiones, la motivación a soñar en grande y el amor que me han dado. Este logro no existiría sin ellos. Es un logro en conjunto.

Agradezco a mis amigos y profesores de la Facultad de Ciencias, en especial a Ricardo, Carlos, Nico, Paoli, Nathan, Paulina y Jorge. Gracias por tantas anécdotas y por el gran ambiente que compartimos a lo largo de nuestros años de estudio.

A Sophia, Jorge, Susie, Marco, Ricardo, Josué, Axel y Luis Heblén por acompañarme en este trayecto y estar conmigo en los momentos buenos y malos. Les debo mucho, amigos míos.

Le agradezco a Andrea por ser mi compañera en este viaje, por ayudarme en la elaboración de este trabajo, por su dedicación en explicarme aspectos del mismo, por motivarme, por su apoyo incondicional y por su cariño.

Agradezco a la Dra. Guillermina Eslava por su dedicación, paciencia, tiempo y guía en este proyecto. Gracias profesora.

Finalmente le agradezco a la UNAM por brindarme la oportunidad de realizar mi licenciatura y por todas las anécdotas y experiencias que he vivido gracias a ella.

Índice general

Algunos métodos de aprendizaje supervisado

por
Gerardo Martínez Mejía

Resumen

Este trabajo expone el desempeño empírico de algunos métodos de aprendizaje supervisado para el caso binario en la variable respuesta, con un conjunto de variables explicativas continuas y discretas. Estos métodos son comparados en términos del error de clasificación en un estudio de cáncer de mama y uno de simulación.

En la primera parte se presentaron los dos conjuntos de datos utilizados. Uno de ellos proviene de un estudio de cáncer de mama; se incluye el significado de las variables presentes, se exponen algunas estadísticas descriptivas y se grafican las 186 observaciones buscando diferenciarlas. Para el conjunto de datos simulados se expone el modelo gráfico de donde provienen, así como su función de densidad y se resume la distribución Gaussiana Condicional en una tabla con respecto al valor de las variables discretas.

En la segunda parte se presentaron los métodos de clasificación supervisada empleados en los dos conjuntos de datos antes mencionados. Se utilizó a la regresión logística, regresión logística a la que se le aplicó *stepwise selection* con criterios AIC y BIC, métodos regularizados como *ridge regression, lasso* y *elastic net* empleados en la regresión logística, análisis discriminante lineal y cuadrático y finalmente, *random forest*.

La última parte estuvo dedicada a las aplicaciones, utilizando el lenguaje de programación **R**, para los dos conjuntos de datos antes mencionados. En cada conjunto se abordó el caso de clasificación utilizando *validation set approach* usando mil repeticiones. Para el primer conjunto de datos se utilizaron las variables sin transformar y se particionó al conjunto en 90 % entrenamiento y 10 % validación y se obtuvo que, a pesar de nuevos métodos algorítmicos como *random forest*, el método que mejor desempeño obtuvo fue la regresión logística a la que se le aplicó *stepwise selection* con criterio BIC con un error de validación global de 33.25 % seguido de la regularización *lasso* y *elastic net* con un error global de 33.29 % y 33.96 %, respectivamente. *Random forest* obtuvo un error de validación de 37.88 %. Los errores de clasificación global tuvieron un rango entre 33-40 %, para la clase 0 un rango entre 26-37 % y para la clase 1, 39-47 %. Los métodos presentaron un rango, para la desviación estándar de la media del error global, entre 0.24-0.37.

Para el estudio de simulación, se generaron conjuntos de entrenamiento y validación de 200 observaciones cada uno con 100 observaciones por clase, considerando 4 variables discretas y 6 continuas. En este caso se observó que el análisis discriminante cuadrático obtuvo el menor error de clasificación global con un 24.77 %, seguido por los métodos de regularización aplicados a la regresión logística con un error entre 26-27 % y la regresión logística a la que se le aplicó *stepwise selection* con criterio BIC con un error de validación global de 26.63 %. En este caso se presentó un rango, para la desviación estándar de la media del error global, entre 0.10-0.12.

Introducción

El avance del poder computacional en el siglo XXI ha hecho que se aprovechen modelos estadísticos del siglo XX como la regresión logística y el análisis discriminante para el análisis y clasificación de datos. El avance computacional ha traído nuevos retos en el área de análisis y predicción de datos, en particular el aprendizaje estadístico automatizado.

En este trabajo se abordaron, para el caso de clasificación, los métodos de aprendizaje supervisado: regresión logística, análisis discriminante y *random forest*, aunado a métodos de regularización para la regresión logística.

El aprendizaje supervisado es utilizado en muchas áreas desde las sociales y económicas hasta las ciencias médicas. En este trabajo se utilizó para predecir el resultado de una cirugía ablativa para el cáncer de mama con la ayuda de variables explicativas. Una aplicación es utilizar a las variables explicativas para predecir el resultado de una variable respuesta que se encuentra dentro del conjunto de datos, a esto se le conoce como *Aprendizaje Supervisado* (Hastie et al., 2009, p.9).
En clasificación supervisada se busca clasificar a una observación en alguna de las K clases conocidas con la ayuda de p variables explicativas y, con los datos disponibles, estimar una regla de decisión que clasifique a las futuras observaciones.

Aunado a los modelos estadísticos clásicos como la regresión logística y análisis discriminante, en este trabajo se expone al método algorítmico *random forest*, propuesto por Leo Breiman.

Este trabajo consiste de 3 capítulos, en el primero se exponen los dos conjuntos de datos utilizados. El primer conjunto proviene de un estudio de cáncer de mama avanzado con 186 observaciones que describe el resultado de una cirugía ablativa: exitoso o intermedio y fracaso. El segundo es el resultado de una simulación de datos de una distribución Gaussiana Condicional con 4 variables binarias, 6 continuas y una variable respuesta con dos clases. En este caso se considera que las variables binarias son independientes de las continuas, por lo tanto para cada celda de cada clase se obtiene la misma distribución Gaussiana.

En el segundo capítulo se abordan los métodos de aprendizaje supervisado, para los métodos regularizados se presenta el problema de optimización que representa *ridge regression, lasso* y *elastic net*, se explica la diferencia entre *ridge regression* y *lasso* puesto que *lasso* tiene la posibilidad de hacer cero a los coeficientes pero *ridge regression* no, se expuso la manera de obtener el valor de λ y, para *elastic net*, de

α. Para el método de *random forest* se describe su algoritmo y se explican algunos parámetros que ocupa. En *stepwise selection* se abordó únicamente el algoritmo de *backward stepwise selection* que comienza con el modelo completo y va eliminando a las variables menos significativas en términos de predicción.
En este capítulo se describe la regla de Bayes que dicta una forma de clasificar a las observaciones dentro de las clases conocidas. Se explica el método de *validation set approach* que se utilizó para medir el poder predictivo de los métodos de aprendizaje usando n repeticiones.
Los métodos abordados aparecen descritos en varios libros que considero fundamentales para el aprendizaje estadístico automatizado como *Computer Age Statistical Inference, Student Edition: Algorithms, Evidence, and Data Science* de Efron & Hastie, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* de Hastie, Tibshirani & Friedman, *An Introduction to Statistical Learning: with Applications in R* de James, Witten, Hastie & Tibshirani, entre otros que pueden ser consultados en la bibliografía de este trabajo.

En el tercer capítulo se aplicaron los métodos de aprendizaje a los dos conjuntos de datos mencionados. Se utilizó *validation set approach* usando mil repeticiones para medir el poder predictivo de cada método. Para cada método se presentaron los errores de validación por clase y global. Para los métodos de regularización y selección de variables, se calculó la media del número de coeficientes presentes en cada modelo y se expuso la metodología utilizada para la selección de los parámetros λ y α. Para *random forest* se buscó la mejor combinación entre *mtry* y *ntree* utilizando al conjunto de entrenamiento y luego, se calcularon los errores de prueba utilizando los mejores valores. Finalmente, se comparan los resultados obtenidos en cada método.

Se utilizó el lenguaje de programación **R** y la paquetería utilizada, así como el código implementado, se exponen en Apéndice B y Apéndice C.

Capítulo 1

Bases de Datos Utilizadas

En este trabajo se utilizan dos conjuntos de datos. El primer conjunto, al cual denotamos por `DatosCancer`, consiste de datos extraídos de un estudio de cáncer de mama avanzado. El segundo conjunto, al cual denotamos por `DatosSim`, está basado en el estudio de simulación hecho por Eslava, G & Pérez, G. (2022) el cual consiste en datos simulados provenientes de una distribución Gaussiana Condicional.

1.1. Estudio de Cáncer de Mama

El conjunto de datos `DatosCancer` ha sido utilizado en varios artículos y la descripción aparece de manera más amplia en Armitage et al. (1969).

El conjunto se compone de una muestra de 186 mujeres que se sometieron a una cirugía ablativa para el cáncer de mama avanzado entre 1958 y 1965 en el Hospital Guy ubicado en Londres. Obteniendo como resultado que en 99 casos el tratamiento fue catalogado como exitoso o intermedio y en 87 casos como un fracaso.

El conjunto de datos cuenta con 9 variables explicativas (6 variables continuas y 3 variables binarias) y una variable respuesta en donde se determinó el resultado de la cirugía. La descripción de cada variable, así como su rango, se encuentra en la Tabla 1.1 (Krzanowski, 1976). En la Tabla 1.2 se presentan algunas estadísticas de las variables continuas. Se observaron desviaciones estándares, por variable, desde 1.09 hasta 677.97. En las medias se tiene un rango entre 3.07 y 819.70.

En la Tabla 1.3 se presentan las correlaciones entre las variables continuas de la Tabla 1.1. Se destaca la correlación global de 0.72 entre la cantidad de Dehydroepiandrosterone en μg cada 24hrs. y la cantidad de Aetiocholanolone en μg cada 24hrs., es decir, entre las variables $\mathtt{x}_5$ y $\mathtt{x}_6$.

Usando un análisis de componentes principales (PCA), donde se consideraron a las variables binarias como numéricas, se graficaron las 186 observaciones buscando distinguirlas dependiendo de la variable respuesta `Y`. Se puede observar en las Figuras 1.1, 1.2, A.1, A.2, A.3 y A.4 una aglomeración de las observaciones sin una clara distinción con respecto a las clases. En la Tabla A.1 se presentan a los eigenvalores

calculados y el porcentaje de varianza explicada por cada uno, se observa que, en el caso de las variables sin estandarizar, los primeros dos componentes principales representan más del 95 % de varianza haciendo ver que, aunque el espacio originalmente es de 10 dimensiones, está contenido en un subespacio de menos dimensiones.

Tabla 1.1: Descripción y rango de las variables presentes en el conjunto de datos `DatosCancer`.

Variable	Descripción	Valores
`Y`	Indicador del resultado de la operación	1= exitosa o intermedia 0= fracaso
$\mathtt{x}_1$	Edad en años en el momento de la mastectomía o cuando se vio por primera vez	[23,69]
$\mathtt{x}_2$	log(tiempo en meses hasta la ablación)	[0,5.65]
$\mathtt{x}_3$	17-hydroxicorticosteroids en mg cada 24hrs.	[1.10,27.30]
$\mathtt{x}_4$	Androsterone en μg cada 24hrs.	[0,1481]
$\mathtt{x}_5$	Dehydroepiandrosterone en μg cada 24hrs.	[20,4434]
$\mathtt{x}_6$	Aetiocholanolone en μg cada 24hrs.	[20,2558]
$\mathtt{x}_7$	Mastectomía	1= afirmativo 0= negativo
$\mathtt{x}_8$	Tipo de ablación	1= andrenalectomy 0= hypophysectomy
$\mathtt{x}_9$	Lesión en el pecho	1= afirmativo 0= negativo

Tabla 1.2: Estadísticas descriptivas de las variables continuas del estudio de cáncer de mama.

	$\mathtt{x}_1$	$\mathtt{x}_2$	$\mathtt{x}_3$	$\mathtt{x}_4$	$\mathtt{x}_5$	$\mathtt{x}_6$
Mín.	23.00	0.00	1.10	0.00	20.00	20.00
Mediana	47.00	3.31	8.80	159.5	615.00	687.00
Media	47.14	3.07	9.66	222.30	819.70	792.80
Máx.	69.00	5.65	27.30	1481.00	4434.00	2558.00
Desv. Est.	8.86	1.09	4.18	247.21	677.97	529.85

Tabla 1.3: Correlación entre las variables continuas del estudio de cáncer de mama presentadas por clase y global.

corr(x_i, x_j)	Global	Clase 1	Clase 0	corr(x_i, x_j)	Global	Clase 1	Clase 0
(x_1, x_2)	-0.13	-0.13	-0.12	(x_2, x_6)	-0.08	-0.06	-0.17
(x_1, x_3)	-0.15	-0.21	-0.10	(x_3, x_4)	0.06	-0.03	0.15
(x_1, x_4)	3e-3	0.15	-0.12	(x_3, x_5)	**0.44**	**0.43**	**0.50**
(x_1, x_5)	**-0.25**	-0.16	**-0.30**	(x_3, x_6)	**0.31**	0.18	**0.43**
(x_1, x_6)	-0.09	-0.02	-0.13	(x_4, x_5)	**0.31**	0.24	**0.35**
(x_2, x_3)	-0.08	0.05	-0.20	(x_4, x_6)	**0.45**	**0.38**	**0.50**
(x_2, x_4)	-0.03	-0.07	-0.02	(x_5, x_6)	**0.72**	**0.60**	**0.76**
(x_2, x_5)	-0.03	-0.08	-0.06				

Nota: Los valores de correlación más altos están destacados en negritas.

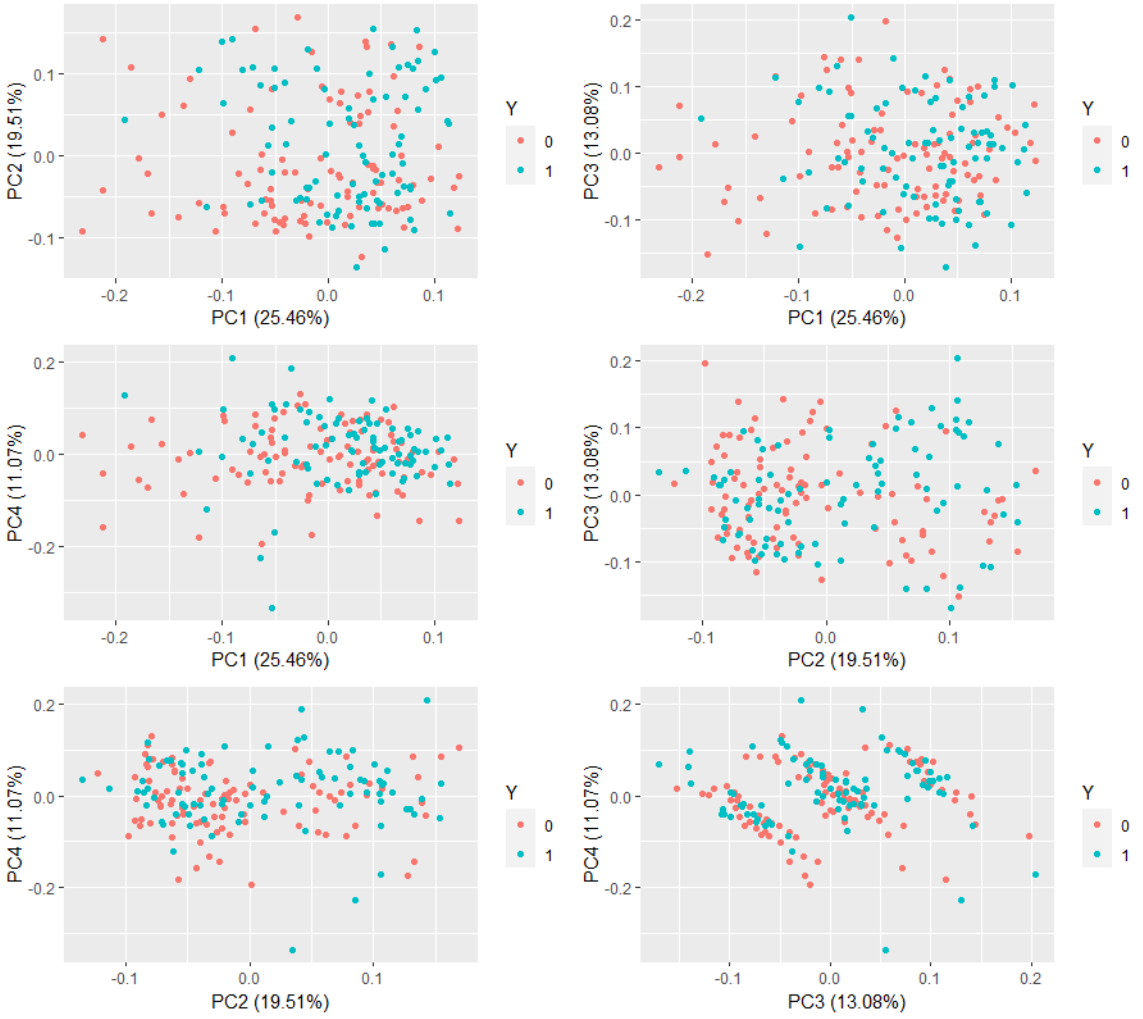

Figura 1.1: Primeros 4 componentes principales para el conjunto de `DatosCancer` con las variables estandarizadas con media cero y varianza uno.

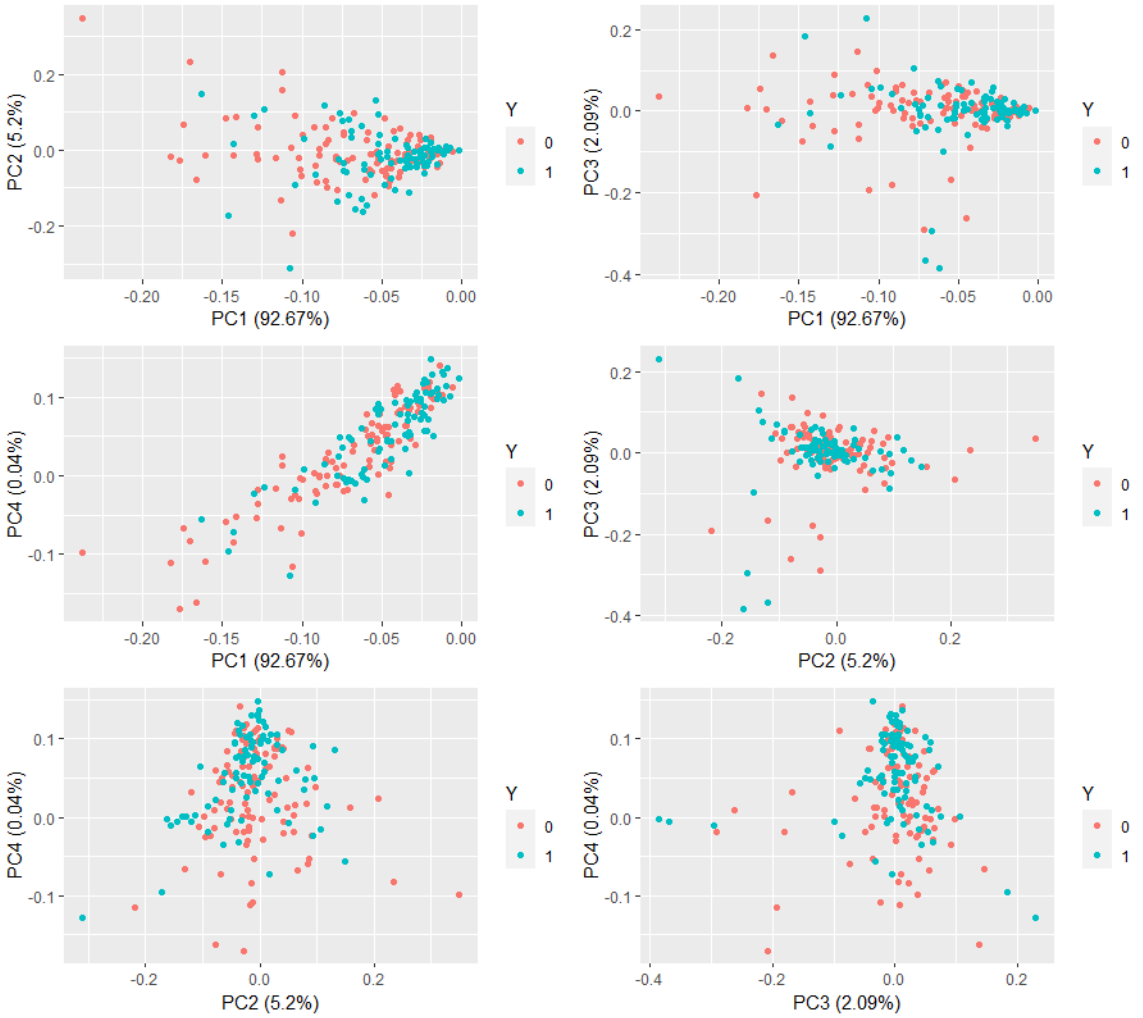

Figura 1.2: Primeros 4 componentes principales para el conjunto de `DatosCancer` con las variables sin estandarizar.

1.2. Datos Simulados

El conjunto de datos `DatosSim` parte del estudio de simulación hecho por Eslava, G. & Pérez, G. (2022) acerca de los modelos gráficos no dirigidos probabilísticos. El código mostrado en el Apéndice C acerca del estudio de simulación de los datos fue proporcionado por los autores de Eslava, G. & Pérez, G. (2022). El conjunto cuenta con 10 variables: 4 binarias y 6 continuas.

$$(\mathbf{i}, \mathbf{x}) = (\mathtt{i}_1, \ldots, \mathtt{i}_4, \mathtt{x}_1, \ldots, \mathtt{x}_6)$$

Con $\mathbf{i} \in \{(0,0,0,0), (0,0,0,1), \ldots, (1,1,1,1)\}$ y $\mathbf{x} \in \mathbb{R}^6$.

Dada una gráfica $G = (V, E)$[1], se define a la factorización por *cliques* como al

[1]Donde V es el conjunto de vértices y E el de aristas.

conjunto de todas las subgráficas maximales[2] $C = \{C_1, \ldots, C_m\}$, donde un *clique* es alguna subgráfica máximal C_i (Anderson, 2009, p.602).
La Figura 1.3 muestra la gráfica asociada a la función Gaussiana Condicional con dos caminos independientes $\mathtt{i}_1 - \cdots - \mathtt{i}_4$ y $\mathtt{x}_1 - \cdots - \mathtt{x}_6$, donde los *cliques* en la gráfica para las variables binarias son: $\mathtt{i}_1 - \mathtt{i}_2, \mathtt{i}_2 - \mathtt{i}_3, \mathtt{i}_3 - \mathtt{i}_4$.
Para calcular las probabilidades de los *cliques*, se define una correlación ρ utilizada para las variables binarias y continuas que depende de la clase $k \in \{0, 1\}$. Se define, para la clase k, a la probabilidad conjunta de que $\mathtt{i}_j = l$ e $\mathtt{i}_h = m$ de la siguiente manera

$$p_k(l, m) = P(\mathtt{i}_j = l, \mathtt{i}_h = m | K = k),$$

con $l, m \in \{0, 1\}$.

Para simular los datos se consideró que $p_k(\mathtt{i}_1, \mathtt{i}_2) = p_k(\mathtt{i}_2, \mathtt{i}_3) = p_k(\mathtt{i}_3, \mathtt{i}_4)$. Para obtener $p_k(1, 1)$ y las demás probabilidades de los *cliques*, basta con definir la probabilidad marginal $p_k(1) = P(\mathtt{i}_j = 1 | K = k) = 0.5$ para $k \in \{0, 1\}$, obteniendo

$$\begin{aligned} p_k(1,1) &= \rho[p_k(1) - p_k(1)^2] + p_k(1)^2 \\ p_k(0,1) &= p_k(1) - p_k(1,1) \\ p_k(1,0) &= p_k(0,1) \\ p_k(0,0) &= 1 - p_k(1,1) - 2p_k(0,1) \end{aligned} \tag{1.1}$$

Para la clase 0 se utilizó $\rho = 0.3$ y para la clase 1 $\rho = -0.3$. Las probabilidades, por clase, son las siguientes

Probabilidad	Clase 0	Clase 1
$p_k(1,1) = p_k(0,0)$	0.325	0.175
$p_k(0,1) = p_k(1,0)$	0.175	0.325

Tabla 1.4: Probabilidades de los *cliques* por clase.

Puesto que en la gráfica de la Figura 1.3 hay dos caminos independientes, las variables discretas $\mathbf{i}$ son marginalmente independientes de las continuas $\mathbf{x}$ y la función de densidad se puede expresar como

$$f_k(\mathbf{i}, \mathbf{x}) = p_k(\mathbf{i}) f_k(\mathbf{x}|\mathbf{i}) = p_k(\mathbf{i}) f_k(\mathbf{x}) \tag{1.2}$$

Donde $f_k(\mathbf{x}|\mathbf{i})$ corresponde a la función de densidad de una $N(\mathbf{0}, \Sigma_k)$ para $k \in \{0, 1\}$ y con $\mathbf{0}$ el vector de ceros de seis dimensiones, Σ_k expresada de la siguiente manera

$$\Sigma_k = \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 & \rho^5 \\ \rho & 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ \rho^2 & \rho & 1 & \rho & \rho^2 & \rho^3 \\ \rho^3 & \rho^2 & \rho & 1 & \rho & \rho^2 \\ \rho^4 & \rho^3 & \rho^2 & \rho & 1 & \rho \\ \rho^5 & \rho^4 & \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}$$

[2]Se define a una subgráfica maximal como a un subconjunto de vértices de V tal que cada dos vértices del subconjunto son adyacentes (Anderson, 2009, p.602).

Con $|\Sigma_k| = -(\rho^2 - 1)^5$ y matriz de concentración

$$\Lambda_k \equiv \Sigma_k^{-1} = \frac{1}{1-\rho^2}\begin{pmatrix} 1 & -\rho & 0 & 0 & 0 & 0 \\ -\rho & 1+\rho^2 & -\rho & 0 & 0 & 0 \\ 0 & -\rho & 1+\rho^2 & -\rho & 0 & 0 \\ 0 & 0 & -\rho & 1+\rho^2 & -\rho & 0 \\ 0 & 0 & 0 & -\rho & 1+\rho^2 & -\rho \\ 0 & 0 & 0 & 0 & -\rho & 1 \end{pmatrix}.$$

Para la clase 0 se utilizó $\rho = 0.3$, las expresiones de Σ_0 y Λ_0 son las siguientes

$$\Sigma_0 = \begin{pmatrix} 1.000 & 0.300 & 0.090 & 0.027 & 0.008 & 0.002 \\ 0.300 & 1.000 & 0.300 & 0.090 & 0.027 & 0.008 \\ 0.090 & 0.300 & 1.000 & 0.300 & 0.090 & 0.027 \\ 0.027 & 0.090 & 0.300 & 1.000 & 0.300 & 0.090 \\ 0.008 & 0.027 & 0.090 & 0.300 & 1.000 & 0.300 \\ 0.002 & 0.008 & 0.027 & 0.090 & 0.300 & 1.000 \end{pmatrix}$$

$$\Lambda_0 = \begin{pmatrix} 1.099 & -0.330 & 0 & 0 & 0 & 0 \\ -0.330 & 1.198 & -0.330 & 0 & 0 & 0 \\ 0 & -0.330 & 1.198 & -0.330 & 0 & 0 \\ 0 & 0 & -0.330 & 1.198 & -0.330 & 0 \\ 0 & 0 & 0 & -0.330 & 1.198 & -0.330 \\ 0 & 0 & 0 & 0 & -0.330 & 1.099 \end{pmatrix}$$

Para la clase 1 se consideró $\rho = -0.3$ con

$$\Sigma_1 = \begin{pmatrix} 1.000 & -0.300 & 0.090 & -0.027 & 0.008 & -0.002 \\ -0.300 & 1.000 & -0.300 & 0.090 & -0.027 & 0.008 \\ 0.090 & -0.300 & 1.000 & -0.300 & 0.090 & -0.027 \\ -0.027 & 0.090 & -0.300 & 1.000 & -0.300 & 0.090 \\ 0.008 & -0.027 & 0.090 & -0.300 & 1.000 & -0.300 \\ -0.002 & 0.008 & -0.027 & 0.090 & -0.300 & 1.000 \end{pmatrix}$$

$$\Lambda_1 = \begin{pmatrix} 1.099 & 0.330 & 0 & 0 & 0 & 0 \\ 0.330 & 1.198 & 0.330 & 0 & 0 & 0 \\ 0 & 0.330 & 1.198 & 0.330 & 0 & 0 \\ 0 & 0 & 0.330 & 1.198 & 0.330 & 0 \\ 0 & 0 & 0 & 0.330 & 1.198 & 0.330 \\ 0 & 0 & 0 & 0 & 0.330 & 1.099 \end{pmatrix}$$

Dados los *cliques* en las variables binarias, la función $p_k(\mathbf{i})$ se puede descomponer en (Anderson, 2009, p.603)

$$p_k(\mathbf{i}) = p_k(\mathtt{i}_1, \mathtt{i}_2, \mathtt{i}_3, \mathtt{i}_4) = \frac{p_k(\mathtt{i}_1, \mathtt{i}_2)p_k(\mathtt{i}_2, \mathtt{i}_3)p_k(\mathtt{i}_3, \mathtt{i}_4)}{p_k(\mathtt{i}_2)p_k(\mathtt{i}_3)} \tag{1.3}$$

Como las variables continuas son independientes de $\mathbf{i}$ y están correlacionadas en cadena: $\mathtt{x}_1 - \mathtt{x}_2 - \cdots - \mathtt{x}_6$, su función de densidad $f_k(\mathbf{x}|\mathbf{i})$ se puede factorizar de la siguiente manera

$$f_k(\mathbf{x}|\mathbf{i}) = f_k(\mathtt{x}_1, \ldots, \mathtt{x}_6|\mathtt{i}_1, \ldots, \mathtt{i}_4) = f_k(\mathtt{x}_1) \prod_{i=2}^{6} f_k(\mathtt{x}_i|\mathtt{x}_{i-1}), \tag{1.4}$$

donde $f_k(\mathtt{x}_i|\mathtt{x}_{i-1})$ corresponde a la función de densidad condicional de la distribución Gaussiana y $f_k(\mathtt{x}_1) \sim N(0,1)$.

De (1.2), (1.3) y (1.4), la función de densidad de la gráfica de la Figura 1.3 es

$$f_k(\mathbf{i}, \mathbf{x}) = \frac{p_k(\mathtt{i}_1, \mathtt{i}_2)p_k(\mathtt{i}_2, \mathtt{i}_3)p_k(\mathtt{i}_3, \mathtt{i}_4)}{p_k(\mathtt{i}_2)p_k(\mathtt{i}_3)} f_k(\mathtt{x}_1) \prod_{i=2}^{6} f_k(\mathtt{x}_i|\mathtt{x}_{i-1}). \tag{1.5}$$

Si se considera a la clase 0 y una observación $(1, 0, 1, 1, 0.5, -0.27, 1, 0.69, 0.2, -0.8)$ entonces

$$\begin{aligned} p_0(i) = p_0(1,0,1,1) = \frac{p_0(1,0)p_0(0,1)p_0(1,1)}{p_0(0)p_0(1)} &= \frac{0.175 \times 0.175 \times 0.325}{0.5 \times 0.5} \\ &= 0.0398 \end{aligned} \tag{1.6}$$

Con $f_0(\mathbf{x}|\mathbf{i})$ expresada de la siguiente manera

$$f_0(\mathbf{x}|\mathbf{i}) = f_0(\mathtt{x}_1)f_0(\mathtt{x}_2|\mathtt{x}_1)f_0(\mathtt{x}_3|\mathtt{x}_2)f_0(\mathtt{x}_4|\mathtt{x}_3)f_0(\mathtt{x}_5|\mathtt{x}_4)f_0(\mathtt{x}_6|\mathtt{x}_5)$$

Donde para la clase 0, $j \in \{2, \ldots, 6\}$ y de (2.73) y (2.74), $f_{\mathtt{x}_j|\mathtt{x}_{j-1}}(x_j)$ está expresada como

$$f_{\mathtt{x}_j|\mathtt{x}_{j-1}}(x_j) = \frac{1}{(2\pi \times 0.91)^{1/2}} \exp\left[-\frac{1}{2 \times 0.91}(x_j - 0.3x_{j-1})^2\right]$$

Entonces

$$\begin{aligned} f_0(x|\mathbf{i}) &= f_0(0.5)f_0(\mathtt{x}_2|0.5)f_0(\mathtt{x}_3|-0.27)f_0(\mathtt{x}_4|1)f_0(\mathtt{x}_5|0.69)f_0(\mathtt{x}_6|0.2) \\ &= 0.6915 \times 0.3222 \times 0.8826 \times 0.6659 \times 0.4969 \times 0.1723 \\ &= 0.0112 \end{aligned} \tag{1.7}$$

Finalmente

$$f_0(i, x) = p_0(i)f_0(x|i) = 0.0398 \times 0.0112 = 0.0004 \tag{1.8}$$

En la Tabla 1.5 se expresan, para $k \in \{0, 1\}$, los valores de $p_k(\mathbf{i})$, μ y Σ_k (1.2) de la función $f_k(\mathbf{i}, \mathbf{x})$ con respecto a las celdas $i \in \{(0,0,0,0), (0,0,0,1), \ldots, (1,1,1,1)\}$. Se observa que en este estudio de simulación, se supuso que la matriz Σ_k es igual entre las celdas de la misma clase y distinta entre las celdas de clases diferentes.

En las Figuras 1.4, 1.5 y de A.5 a A.10 se observan a los componentes principales de un conjunto de datos con 50 observaciones por clase, considerando a todas las variables como numéricas y las características antes mencionadas. La primeras 3

figuras corresponden a las variables estandarizadas y las últimas tres a las variables sin estandarizar. En la Tabla A.4 se presentan a los eigenvalores calculados y en la Tabla A.5 y Tabla A.6 a los eigenvectores.

	Clase 0			Clase 1		
i	$p_0(\mathbf{i})$	μ	Σ	$p_1(\mathbf{i})$	μ	Σ
(0,0,0,0)	0.1373	$\mathbf{0}$	$\rho = 0.3$	0.0214	$\mathbf{0}$	$\rho = -0.3$
(1,0,0,0)	0.0739	$\mathbf{0}$	$\rho = 0.3$	0.0398	$\mathbf{0}$	$\rho = -0.3$
(0,1,0,0)	0.0398	$\mathbf{0}$	$\rho = 0.3$	0.0739	$\mathbf{0}$	$\rho = -0.3$
(0,0,1,0)	0.0398	$\mathbf{0}$	$\rho = 0.3$	0.0739	$\mathbf{0}$	$\rho = -0.3$
(0,0,0,1)	0.0739	$\mathbf{0}$	$\rho = 0.3$	0.0398	$\mathbf{0}$	$\rho = -0.3$
(0,1,1,0)	0.0398	$\mathbf{0}$	$\rho = 0.3$	0.0739	$\mathbf{0}$	$\rho = -0.3$
(1,0,0,1)	0.0398	$\mathbf{0}$	$\rho = 0.3$	0.0739	$\mathbf{0}$	$\rho = -0.3$
(0,0,1,1)	0.0739	$\mathbf{0}$	$\rho = 0.3$	0.0398	$\mathbf{0}$	$\rho = -0.3$
(1,1,0,0)	0.0739	$\mathbf{0}$	$\rho = 0.3$	0.0398	$\mathbf{0}$	$\rho = -0.3$
(0,1,0,1)	0.0214	$\mathbf{0}$	$\rho = 0.3$	0.1373	$\mathbf{0}$	$\rho = -0.3$
(1,0,1,0)	0.0214	$\mathbf{0}$	$\rho = 0.3$	0.1373	$\mathbf{0}$	$\rho = -0.3$
(1,1,1,0)	0.0739	$\mathbf{0}$	$\rho = 0.3$	0.0398	$\mathbf{0}$	$\rho = -0.3$
(0,1,1,1)	0.0739	$\mathbf{0}$	$\rho = 0.3$	0.0398	$\mathbf{0}$	$\rho = -0.3$
(1,0,1,1)	0.0398	$\mathbf{0}$	$\rho = 0.3$	0.0739	$\mathbf{0}$	$\rho = -0.3$
(1,1,0,1)	0.0398	$\mathbf{0}$	$\rho = 0.3$	0.0739	$\mathbf{0}$	$\rho = -0.3$
(1,1,1,1)	0.1373	$\mathbf{0}$	$\rho = 0.3$	0.0214	$\mathbf{0}$	$\rho = -0.3$

Tabla 1.5: Descripción de $p_k(\mathbf{i})$, μ y Σ con respecto a la clases y las celdas $\mathbf{i}$. La matriz Σ está expuesta en (1.2). Se observa que la matriz Σ es diferente entre clases e igual entre las celdas de la misma clase.

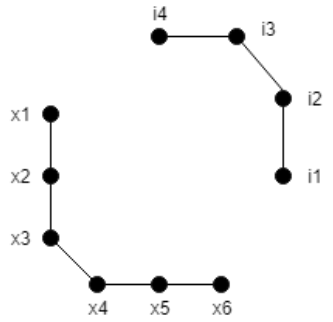

Figura 1.3: Gráfica asociada a la función Gaussiana Condicional $f_k(\mathbf{i}, \mathbf{x}) = p_k(\mathbf{i}) f_k(\mathbf{x}|\mathbf{i})$ para la simulación del conjunto de datos `DatosSim`.

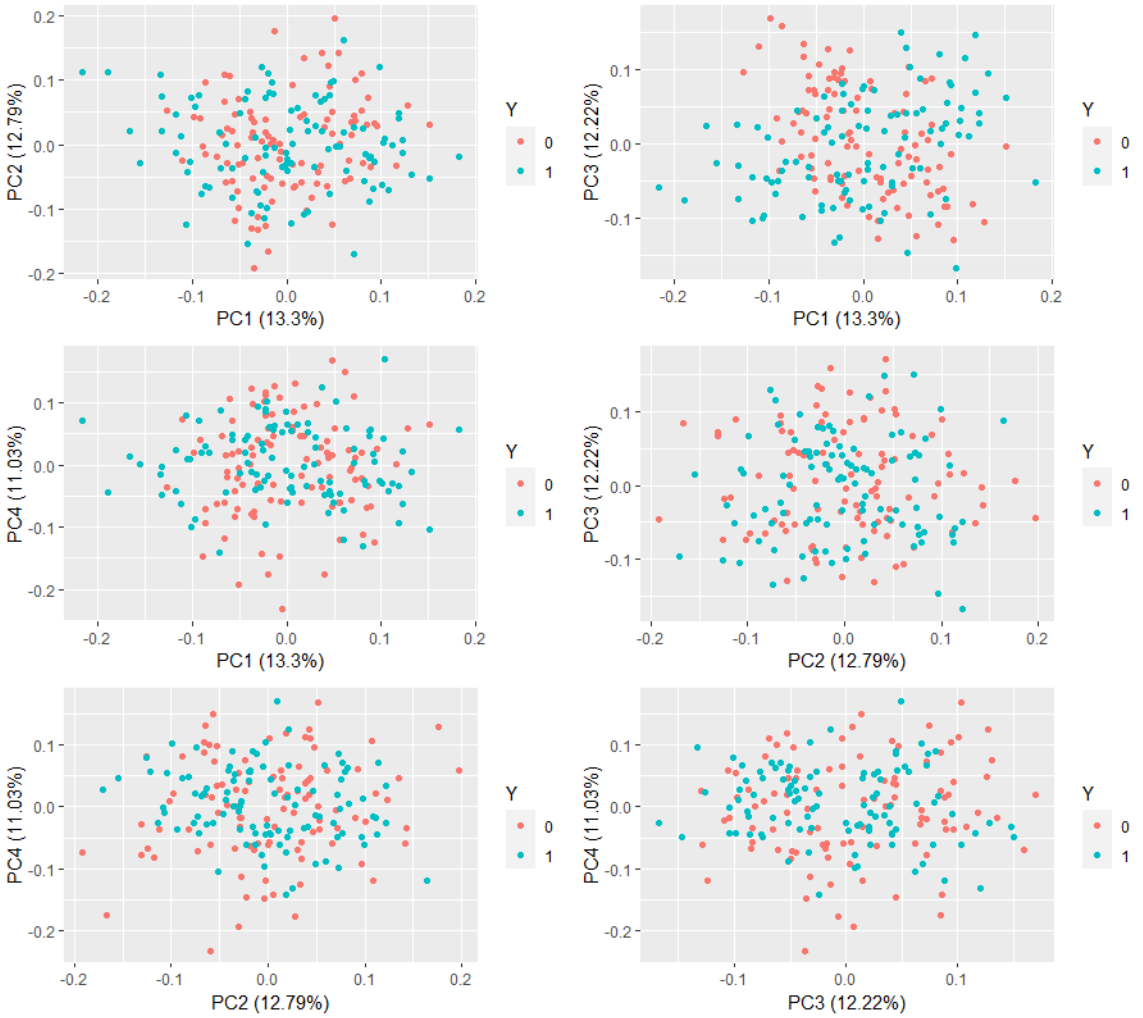

Figura 1.4: Primeros 4 componentes principales para el conjunto de datos `DatosSim` con las variables estandarizadas con media cero y varianza uno.

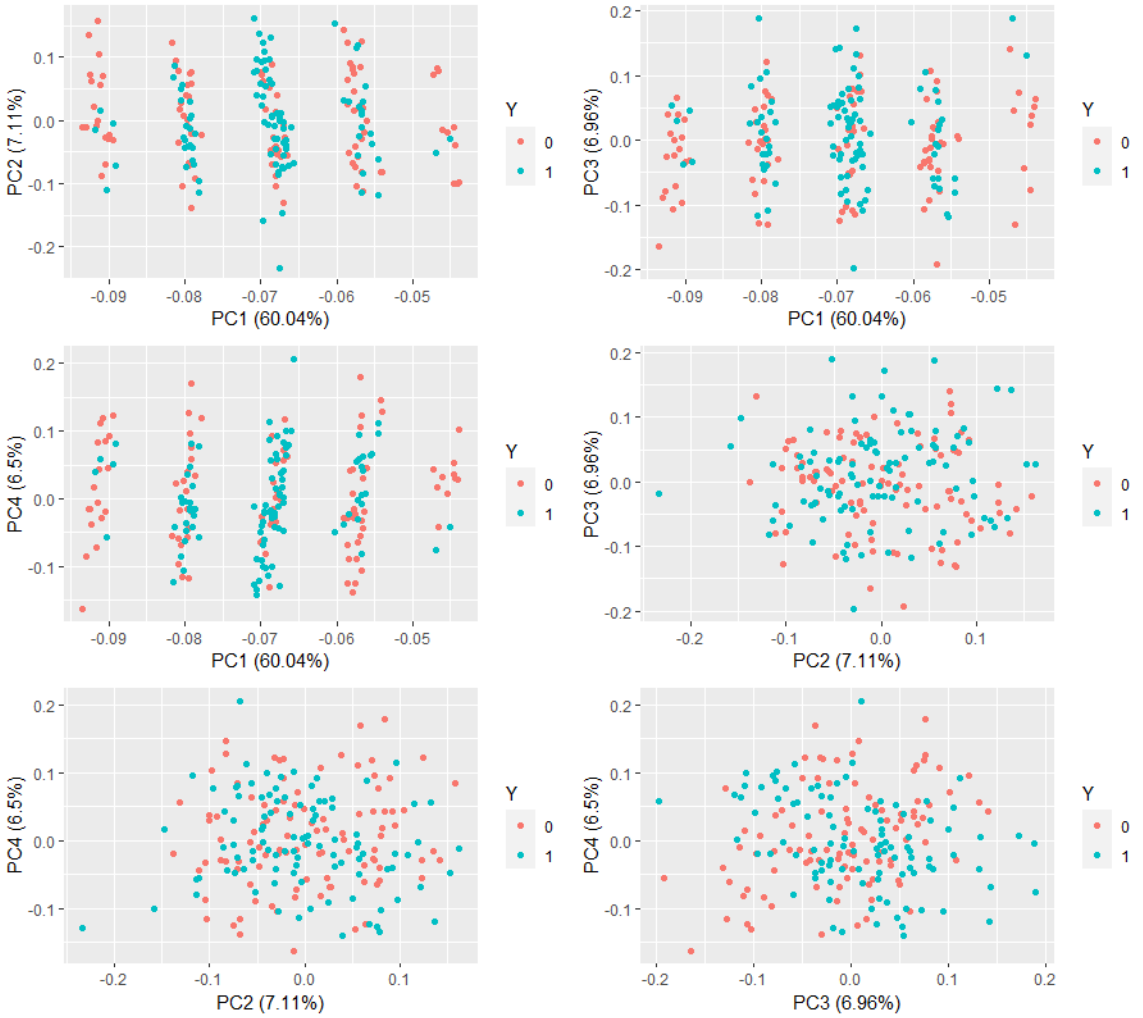

Figura 1.5: Primeros 4 componentes principales para el conjunto de datos `DatosSim` con las variables sin estandarizar.

En este capítulo se presentaron los dos conjuntos de datos que se utilizaron a lo largo de este trabajo. En el siguiente capítulo se describen los métodos de clasificación supervisada contenidos en esta tesis.

Capítulo 2

Algunos Métodos de Clasificación Supervisada

2.1. Regla de Bayes

Suponga que se tienen n observaciones, un vector $\mathbf{x}$ de p variables explicativas y K clases. La expresión de la probabilidad *a priori* de la clase k es la siguiente

$$\pi_k = P(\mathtt{Y} = k) \tag{2.1}$$

y la probabilidad *a posteriori* es

$$P(\mathtt{Y} = k|\mathbf{x}) \tag{2.2}$$

La *Regla de Clasificación de Bayes* consiste en clasificar a una observación x en la clase k que tenga mayor probabilidad *a posteriori* (Venables & Ripley, 2002, p.333). Para el caso binario, la regla de Bayes clasifica a la observación x en la clase 1 si

$$P(\mathtt{Y} = 1|\mathbf{x} = x) > P(\mathtt{Y} = 0|\mathbf{x} = x) \tag{2.3}$$

Si se asume que $\mathbf{x}$ tiene una función de densidad $f_k(\mathbf{x})$ para la clase k, la regla de Bayes (2.3) es equivalente a clasificar a una observación x en la clase 1 si

$$\log\left(\frac{f_1(x)}{f_0(x)}\right) - \log\left(\frac{\pi_0}{\pi_1}\right) > 0 \tag{2.4}$$

La regla de Bayes minimiza la probabilidad de cometer una clasificación errónea $P(e)$ expresada como

$$P(e) = \pi_0 P(1|0) + \pi_1 P(0|1)$$

Donde $P(m|n)$ denota la probabilidad de asignar a una observación de la clase n a la clase m.

2.2. Evaluación y Selección de Modelos

Dado un conjunto de observaciones, hay dos aspectos importantes a considerar: (Hastie et al., 2009, p.222)

- Construcción de una regla de decisión
  Se ajustan diferentes modelos con el propósito de definir una regla de decisión con el mejor modelo.
- Estimación del poder predictivo
  Una vez que se escogió al modelo final, se estiman los errores de predicción en un conjunto de datos nuevo.

En los métodos de aprendizaje supervisado se pueden distinguir dos tipos de errores de predicción:

- Error de entrenamiento, *training error*.
- Error de prueba o de validación, *test or generalization error*.

El error de entrenamiento se calcula con las observaciones del conjunto de entrenamiento y el error de prueba con las del conjunto de prueba o validación.

2.2.1. *Train, Validation and Test Set*

Si se tiene una cantidad significativa de observaciones, la mejor manera de construir una regla de decisión y estimar su poder predictivo es dividir al conjunto de datos original en tres conjuntos disjuntos: conjunto de entrenamiento, conjunto de validación y conjunto de prueba (Hastie et al., 2009, p.222).
El conjunto de entrenamiento es utilizado para ajustar el modelo, el conjunto de validación para calibrar los hiperparámetros del modelo y el conjunto de prueba emula nuevas observaciones con las que se mide el poder predictivo del modelo seleccionado.
En este trabajo se particionaron a los conjuntos de datos en dos: conjunto de entrenamiento y conjunto de prueba.

Figura 2.1: Particiones del conjunto de datos original en conjuntos de entrenamiento, validación y prueba.
Fuente: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* p.222.

En este capítulo se ajustaron varios modelos y se expone al error aparente obtenido. El error aparente consiste en el error de entrenamiento utilizando a todo el conjunto de datos como conjunto de entrenamiento. Para la regresión logística,

está expresado de la siguiente manera

$$\text{error}_{\text{aparente}} = \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \widehat{y}_i)$$

Con n el número de observaciones y $\widehat{y}_i$ el valor estimado de la variable respuesta en la i-ésima observación.

2.2.2. *Validation Set Approach*

En el Capítulo 3 se utilizó *validation set approach* para medir el poder predictivo de los modelos. Este método consiste en particionar al conjunto de datos de manera aleatoria en dos: conjunto de entrenamiento y conjunto de validación. Se realiza el aprendizaje con el conjunto de entrenamiento y se prueba con el de validación. Este proceso se repite $B \geq 1$ veces (James et al., 2013, p.198).

2.3. Regresión Logística

2.3.1. Regresión Logística Simple

El modelo de *Regresión Logística Simple* se utiliza cuando se cuenta con solo una variable explicativa. Dada una variable respuesta $\mathtt{Y}$ y una explicativa $\mathtt{x}$, se define al modelo como

$$P(\mathtt{Y} = 1|\mathtt{x} = x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \tag{2.5}$$

donde β_0 y β_1 son parámetros desconocidos y $\mathtt{Y}; \mathtt{x} \sim \textit{Bernoulli}(P(\mathtt{Y} = 1|\mathtt{x} = x))$.

Dado que el modelo de regresión logística posee un rango entre 0 y 1, se puede hacer una clasificación de $\mathtt{Y}$ dado $\mathtt{x} = x$. Con base a la regla de Bayes se clasifica dentro de la clase 1 si $P(\mathtt{Y} = 1|\mathtt{x} = x) \geq 0.5$ y clase 0 en otro caso, aunque el valor de corte podría ser un hiperparámetro a calibrar.

La función *logit* está definida de la siguiente forma

$$\textit{logit}(x) = \log\left(\frac{x}{1-x}\right)$$

Que al usarla en el modelo de regresión logística (2.5) se obtiene (Hosmer et al., 2013, p.6)

$$\textit{logit}[E(\mathtt{Y}|\mathtt{x})] = \textit{logit}[P(\mathtt{Y} = 1|\mathtt{x} = x)] = \log\left(\frac{P(\mathtt{Y} = 1|\mathtt{x} = x)}{1 - P(\mathtt{Y} = 1|\mathtt{x} = x)}\right) = \beta_0 + \beta_1 x \tag{2.6}$$

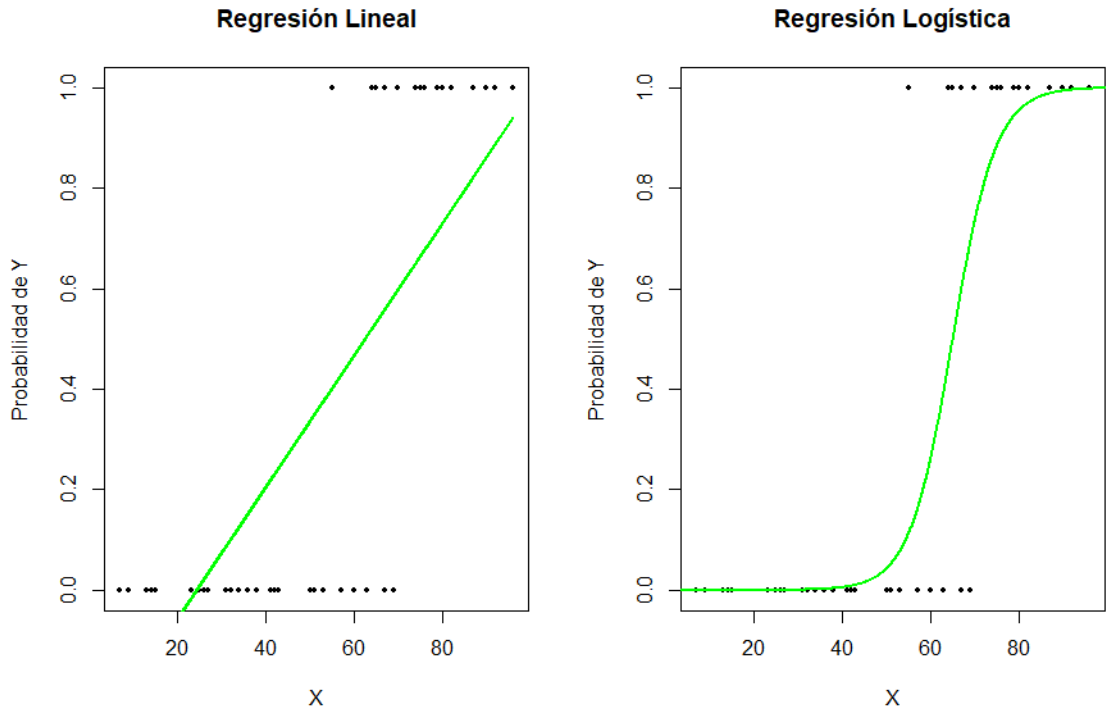

Figura 2.2: Diferencias entre la regresión lineal y regresión logística aplicadas a un conjunto de datos con variable respuesta binaria. Se observa que en la regresión lineal hay observaciones con probabilidades menores a 0 ($X < 20$) y por el contrario, en la regresión logística las probabilidades se encuentran entre 0 y 1.

Estimación y ajuste del modelo

Suponga que se tienen n pares de observaciones independientes (y_i, x_i) donde y_i corresponde al valor de la variable respuesta $\mathtt{Y}$ y x_i al valor correspondiente a la variable explicativa $\mathtt{x}$ en la i-ésima observación. Suponga que $\mathtt{Y}$=1 en n_1 casos y $\mathtt{Y}$=0 en n_2 casos.
Al ajustar un modelo de regresión logística a los datos, se obtienen los parámetros desconocidos β_0 y β_1. Se estiman los parámetros utilizando el método de *Máxima Verosimilitud*. Aplicando la función de *verosimilitud* a la distribución de $\mathtt{Y}$; $\mathtt{x}$ se obtiene lo siguiente

$$\begin{aligned} L(\beta_0, \beta_1) &= \prod_{i=1}^{n_1} [P(\mathtt{Y}=1|\mathtt{x}=x_i)]^{y_i} \prod_{i=1}^{n_2} [1 - P(\mathtt{Y}=1|\mathtt{x}=x_i)]^{1-y_i} \\ &= \prod_{i=1}^{n} [P(\mathtt{Y}=1|\mathtt{x}=x_i)]^{y_i} [1 - P(\mathtt{Y}=1|\mathtt{x}=x_i)]^{1-y_i} \end{aligned} \tag{2.7}$$

Para poder estimar los valores de β_0 y β_1 que maximicen a la ecuación (2.7) se considera a la transformación logaritmo aplicada a la verosimilitud (*logverosimilitud*),

denotada de la siguiente manera

$$\begin{aligned} l(\beta_0, \beta_1) &= \sum_{i=1}^{n} y_i \log[P(\mathtt{Y} = 1|\mathtt{x} = x_i)] + (1 - y_i) \log[1 - P(\mathtt{Y} = 1|\mathtt{x} = x_i)] \\ &= \sum_{i=1}^{n} y_i(\beta_0 + \beta_1 x_i) - \log[1 + \exp(\beta_0 + \beta_1 x_i)] \end{aligned} \tag{2.8}$$

Derivando e igualando a cero la ecuación (2.8) se obtiene el siguiente sistema de ecuaciones

$$\begin{cases} \frac{\partial l}{\partial \beta_0} = \sum_{i=1}^{n} y_i - P(\mathtt{Y} = 1|\mathtt{x} = x_i) = 0 \\ \frac{\partial l}{\partial \beta_1} = \sum_{i=1}^{n} x_i[y_i - P(\mathtt{Y} = 1|\mathtt{x} = x_i)] = 0 \end{cases} \tag{2.9}$$

El sistema de ecuaciones se resuelve de manera numérica utilizando el algoritmo Netwon-Raphson expresado en *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* página 120.
Para asegurar que los valores (β_0, β_1) que resuelven el sistema de ecuaciones correspondan a un punto máximo, se tiene que calcular la matriz Hessiana denotada por

$$H_l = \begin{pmatrix} \frac{\partial^2 l}{\partial \beta_0^2} & \frac{\partial^2 l}{\partial \beta_1 \partial \beta_0} \\ \frac{\partial^2 l}{\partial \beta_1 \partial \beta_0} & \frac{\partial^2 l}{\partial \beta_1^2} \end{pmatrix} \tag{2.10}$$

Y luego, verificar que $|H_l(\beta_0, \beta_1)| > 0$ y $\frac{\partial^2 l}{\partial \beta_0^2}(\beta_0, \beta_1) < 0$.
Los valores de (β_0, β_1) que resuelven el sistema de ecuaciones y cumplen con ser un punto máximo son denominados como *estimadores máximos verosímiles* y se denotan por $(\widehat{\beta}_0, \widehat{\beta}_1)$.
Conociendo los valores de los parámetros, se pueden obtener las probabilidades *a posteriori* estimadas

$$\widehat{P}(\mathtt{Y} = 1|\mathtt{x} = x) = \frac{\exp(\widehat{\beta}_0 + \widehat{\beta}_1 x)}{1 + \exp(\widehat{\beta}_0 + \widehat{\beta}_1 x)} \tag{2.11}$$

De la primera ecuación del sistema de ecuaciones (2.9) se obtiene

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \widehat{P}(\mathtt{Y} = 1|\mathtt{x} = x_i)$$

Exponiendo que la suma de los valores observados es igual a la suma de los valores esperados.

Se ajustó una regresión logística al conjunto de datos `DatosCancer` utilizando a la variable $\mathtt{x}_1$ como única variable explicativa. La Tabla 2.1 muestra los parámetros obtenidos, donde las probabilidades estimadas están dadas por

$$\widehat{P}(\mathtt{Y} = 1|\mathtt{x}_1 = x) = \frac{\exp(-0.7558 + 0.0133x)}{1 + \exp(-0.7558 + 0.0133x)} \tag{2.12}$$

Tabla 2.1: Coeficientes estimados para el conjunto de datos `DatosCancer` utilizando a $\mathtt{x}_1$ como única variable explicativa.

	$\widehat{\beta}$	Desv. Est.	z *value*	Pr(>|z|)
Intercepto	-0.7558	0.8031	-0.94	0.3467
$\mathtt{x}_1$	0.0133	0.0167	0.79	0.4271

Y *logit* estimado

$$logit[\widehat{P}(\mathtt{Y}=1|\mathtt{x}_1=x)] = -0.7558 + 0.0133x \tag{2.13}$$

En la siguiente sección se describen algunas pruebas sobre la significancia estadística del modelo así como algunos intervalos de confianza, si bien ambos rubros no son contemplados al medir el poder predictivo de la regresión logística, se presentan como una extensión de conocimientos.

Pruebas sobre la significancia estadística del modelo

Una vez que se estimaron los parámetros del modelo, se busca comprobar si todos los parámetros están significativamente relacionados con la variable respuesta. La idea es comparar los valores observados de la variable respuesta con valores predichos obtenidos de los modelos con y sin la variable en cuestión (Hosmer et al., 2013, p.12). La comparación entre las verosimilitudes de modelos anidados o submodelos está basada en la *razón de verosimilitud* siguiente

$$D = -2\log\left[\frac{L(\widehat{\beta}_0, \widehat{\beta}^k)}{L(\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p)}\right] \tag{2.14}$$

Donde $(\widehat{\beta}_0, \widehat{\beta}^k)$ es un subconjunto de $(\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p)$.
Aplicando la ecuación (2.8) a (2.14) se obtiene (Hosmer et al., 2013, p.12)

$$D = -2\sum_{i=1}^{n} y_i \log\left(\frac{\widehat{P}(\mathtt{Y}=1|\mathtt{x}=x_i)}{y_i}\right) + (1-y_i)\log\left(\frac{1-\widehat{P}(\mathtt{Y}=1|\mathtt{x}=x_i)}{1-y_i}\right) \tag{2.15}$$

La ecuación anterior es conocida como *devianza*. Para evaluar la significancia estadística de una variable explicativa se calcula la diferencia entre las devianzas obtenidas del modelo sin la variable y el modelo con la variable. Puesto que en ambas devianzas se tiene la verosimilitud del modelo saturado y apoyándose en las propiedades de la función logaritmo, esta diferencia, para el modelo de regresión logística simple, se puede escribir de la siguiente manera

$$\begin{aligned} G &= D(\text{modelo sin } \widehat{\beta}_1) - D(\text{modelo con } \widehat{\beta}_1) \\ &= -2\log\left[\frac{L(\widehat{\beta}_0)}{L(\widehat{\beta}_0, \widehat{\beta}_1)}\right] \end{aligned} \tag{2.16}$$

Dando como resultado que G es

$$\begin{aligned} G =& 2\left\{\sum_{i=1}^{n}[y_i \log[\widehat{P}(\mathtt{Y}=1|\mathtt{x}=x_i)] + (1-y_i)\log\{1-\widehat{P}(\mathtt{Y}=1|\mathtt{x}=x_i)\}]\right. \\ & \left. - \left[\sum_{i=1}^{n} y_i \log\left(\sum_{i=1}^{n} y_i\right) + \sum_{i=1}^{n}(1-y_i)\log\left(\sum_{i=1}^{n}(1-y_i)\right) - n\log(n)\right]\right\} \end{aligned} \tag{2.17}$$

Al suponer que $\beta_1 = 0$, G tiene distribución ji-cuadrada con un grado de libertad. La primera parte de la ecuación (2.17) corresponde al valor de la logverosimilitud.
El valor de G para el modelo expuesto en la Tabla 2.1 es el siguiente

$$\begin{aligned} G &= 2\{-128.2209 - [99\log(99) + 87\log(87) - 186\log(186)]\} \\ &= 2[-128.2209 - (-128.5380)] \\ &= 0.63 \end{aligned}$$

Una vez obteniendo el valor, se procede a calcular el *p-value* asociado a la prueba de la significancia de $\mathtt{x}_1$ sobre la variable respuesta $\mathtt{Y}$, dando como resultado

$$P[\chi^2(1) > 0.63] = 0.42 > 0.05$$

Finalmente, se encontró evidencia estadística que la variable $\mathtt{x}_1$ no es significativa para modelar la $E(\mathtt{Y})$ con un nivel de significancia $\alpha = 0.05$.

Prueba Wald

Dentro de la Tabla 2.1 se encuentra la prueba Wald que determina la significancia estadística de los coeficientes estimados y consiste en el siguiente cociente

$$W = \frac{\widehat{\beta}_1}{\widehat{\mathrm{SE}}(\widehat{\beta}_1)} = \frac{0.0133}{0.0167} = 0.79 \tag{2.18}$$

Al calcular el *p-value* de dos colas $P(|z| > 0.79) = 0.4271$, con z una variable aleatoria con distribución normal estándar, se concluye, con un nivel de significancia $\alpha = 0.05$, que el efecto de la variable $\mathtt{x}_1$ en la variable respuesta $\mathtt{Y}$ no es estadísticamente significativo.

Intervalos de Confianza

Coeficientes del modelo

Normalmente al crear los intervalos de confianza para los coeficientes de un modelo se utiliza la prueba Wald. Un intervalo de $100(1-\alpha)\,\%$ de confianza para el coeficiente estimado $\widehat{\beta}_1$ tiene los siguientes extremos

$$\widehat{\beta}_1 \pm z_{1-\alpha/2}\widehat{\mathrm{SE}}(\widehat{\beta}_1)$$

El intervalo de 95 % de confianza para $\widehat{\beta}_1$ en la Tabla 2.1 tiene los siguientes extremos

$$0.0133 \pm 1.96 \times 0.0167$$

Dando como resultado el intervalo (-0.0194,0.0460). Para calcular un intervalo de $100(1-\alpha)\,\%$ de $\widehat{\beta}_0$ se procede análogamente, obteniendo los siguientes extremos

$$\widehat{\beta}_0 \pm z_{1-\alpha/2}\widehat{\mathrm{SE}}(\widehat{\beta}_0)$$

Varianza del *logit*

El estimador de la varianza del *logit* se calcula de la siguiente manera

$$\widehat{\mathrm{Var}}\{logit[\widehat{P}(\mathtt{Y}=1|\mathtt{x}=x)]\} = \widehat{\mathrm{Var}}(\widehat{\beta}_0) + x^2\widehat{\mathrm{Var}}(\widehat{\beta}_1) + 2x\widehat{\mathrm{Cov}}(\widehat{\beta}_0,\widehat{\beta}_1) \tag{2.19}$$

Usando a (2.19) se define a los puntos extremos del intervalo de confianza como

$$logit[\widehat{P}(\mathtt{Y}=1|\mathtt{x}=x)] \pm z_{1-\alpha/2}\widehat{\mathrm{SE}}\{logit[\widehat{P}(\mathtt{Y}=1|\mathtt{x}=x)]\}$$

Donde $\widehat{\mathrm{SE}}\{logit[\widehat{P}(\mathtt{Y}=1|\mathtt{x}=x)]\}$ es la raíz positiva de la ecuación (2.19).

Se ajustó un modelo de regresión logística al conjunto de datos `DatosCancer` con la variable explicativa binaria $\mathtt{x}_7$. En la Tabla 2.2 se observan los coeficientes estimados del modelo. Al considerar una variable explicativa binaria, el *logit* del modelo se define de la siguiente manera

$$logit[\widehat{P}(\mathtt{Y}=1|\mathtt{x}_7)] = 0.3151 - 0.6797\mathtt{x}_7 = \begin{cases} 0.3151 & \mathtt{x}_7 = 0 \\ -0.3646 & \mathtt{x}_7 = 1 \end{cases} \tag{2.20}$$

En la Subsección 2.3.2 se abordará el caso cuando la variable explicativa sea categórica.

Tabla 2.2: Coeficientes estimados para el conjunto de datos `DatosCancer` utilizando a $\mathtt{x}_7$ como única variable explicativa. Se exponen los coeficientes para el caso $\mathtt{x}_7 = 0$ y $\mathtt{x}_7 = 1$, cuando $\mathtt{x}_7 = 0$ su coeficiente es igual a 0.

	$\widehat{\beta}$	Desv. Est.	z *value*	Pr(>|z|)
Intercepto	0.3151	0.2531	1.24	0.2132
$\mathtt{x}_7 = 1$	-0.6797	0.3130	-2.17	0.0299

Se puede calcular el intervalo de 95 % de confianza para el *logit* de la Tabla 2.2. La matriz de covarianzas estimadas de los estimadores de la Tabla 2.2 es

$$\begin{array}{c} \\ \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{array} \begin{array}{c} \begin{array}{cc} \widehat{\beta}_0 & \widehat{\beta}_1 \end{array} \\ \begin{pmatrix} 0.06 & -0.06 \\ -0.06 & 0.10 \end{pmatrix} \end{array} \tag{2.21}$$

El estimador de la varianza del *logit* es el siguiente

$$\widehat{\text{Var}}\{logit[\widehat{P}(\mathtt{Y}=1|\mathtt{x}_7=1)]\} = 0.06 + 1^2 \times 0.10 + 2 \times 1 \times -0.06 = 0.04$$

Finalmente, los valores extremos del intervalo de 95 % del *logit* son

$$-0.3646 \pm 1.96 \times 0.2$$

El *logit* estimado su intervalo de confianza inducen al intervalo de confianza de $\widehat{P}(\mathtt{Y} = 1|\mathtt{x} = x)$. El intervalo de $100(1-\alpha)$ % de confianza tiene como extremos a los siguientes valores

$$\frac{\exp(logit[\widehat{P}(\mathtt{Y}=1|\mathtt{x}=x)] \pm z_{1-\alpha/2}\widehat{\text{SE}}\{logit[\widehat{P}(\mathtt{Y}=1|\mathtt{x}=x)]\})}{1+\exp(logit[\widehat{P}(\mathtt{Y}=1|\mathtt{x}=x)] \pm z_{1-\alpha/2}\widehat{\text{SE}}\{logit[\widehat{P}(\mathtt{Y}=1|\mathtt{x}=x)]\})}$$

Retomando el modelo expuesto en la Tabla 2.2, cuando $\mathtt{x}_7 = 1$, el intervalo de 95 % de confianza de $\widehat{P}(\mathtt{Y} = 1|\mathtt{x}_7 = 1)$ es

$$\left(\frac{\exp(-0.75)}{1+\exp(-0.75)}, \frac{\exp(0.02)}{1+\exp(0.02)}\right) = (0.32, 0.50)$$

Clasificación

Suponga el modelo expresado en la ecuación (2.12) y una observación $x_1 = 50$. La probabilidad *a posteriori* estimada sería

$$\widehat{P}(\mathtt{Y}=1|\mathtt{x}_1=50) = \frac{\exp(-0.7558 + 0.0133 \times 50)}{1+\exp(-0.7558 + 0.0133 \times 50)} \approx 0.47$$

Lo que implica que $\widehat{P}(\mathtt{Y} = 0|\mathtt{x}_1 = 50) \approx 0.53$ y utilizando la regla de clasificación de Bayes, se clasificará a la observación $\mathtt{x}_1 = 50$ en clase 0.

2.3.2. Regresión Logística Múltiple

Suponga una colección de p variables explicativas agrupadas en un vector $\mathbf{x}^\top = (\mathtt{x}_1, \ldots, \mathtt{x}_p)$, una variable respuesta binaria $\mathtt{Y}$ y n observaciones independientes. Se define al modelo de *Regresión Logística Múltiple* como

$$P(\mathtt{Y}=1|\mathbf{x}) = \frac{\exp(\beta_0 + \beta_1\mathtt{x}_1 + \cdots + \beta_p\mathtt{x}_p)}{1+\exp(\beta_0 + \beta_1\mathtt{x}_1 + \cdots + \beta_p\mathtt{x}_p)} \tag{2.22}$$

El *logit* de la regresión logística múltiple es

$$logit[P(\mathtt{Y}=1|\mathbf{x})] = \log\left(\frac{P(\mathtt{Y}=1|\mathbf{x})}{1-P(\mathtt{Y}=1|\mathbf{x})}\right) = \beta_0 + \beta_1\mathtt{x}_1 + \cdots + \beta_p\mathtt{x}_p \tag{2.23}$$

Al tener más de una variable explicativa puede que alguna(s) de ellas sean categóricas. Estas variables no deben de ser tratadas como numéricas o continuas. Un ejemplo sería

Tabla 2.3: Comportamiento de las variables *dummy*, D_1 y D_2, con respecto al valor de la variable `nacionalidad`.

`nacionalidad`	D_1	D_2
mexicana	0	0
brasileña	1	0
española	0	1

si en algún modelo se toma la variable `nacionalidad` con posibles valores *mexicana*, *española* y *brasileña*. En este caso se crean 2 variables *dummy* denotadas por D_i, las cuales responderían de la siguiente manera: si la nacionalidad es *mexicana* entonces $D_1 = D_2 = 0$, si la nacionalidad es *brasileña* entonces $D_1 = 1$ y $D_2 = 0$ y si la nacionalidad es *española* entonces $D_1 = 0$ y $D_2 = 1$. En la Tabla 2.3 se resume el comportamiento de estas variables *dummy*.

En general, si se tiene una variable explicativa categórica con k niveles, se crearán $k-1$ variables *dummy*.

Suponga que se tienen p variables explicativas, una variable respuesta binaria $\mathtt{Y}$ y la variable x_j categórica con k niveles. Las $k-1$ variables *dummy* se denotarán como D_{jm} y sus coeficientes como β_{jm} con $m = 1, \ldots, k-1$. Finalmente, el *logit* de este modelo sería (Hosmer et al., 2013, p.36)

$$logit[P(\mathtt{Y}=1|\mathbf{x})] = \beta_0 + \beta_1 \mathtt{x}_1 + \cdots + \sum_{m=1}^{k-1} \beta_{jm} D_{jm} + \beta_p \mathtt{x}_p$$

Estimación y ajuste del modelo

Suponga que se tienen p variables explicativas y n pares de observaciones (x_i, y_i) con $i = 1, \ldots, n$ y $x_i = (x_{i1}, \ldots, x_{ip})$. Se busca estimar al parámetro β_0 y al vector $\beta^\top = (\beta_1, \ldots, \beta_p)$ usando máxima verosimilitud. La función de verosimilitud es análoga a la vista en (2.7) y nuevamente se emplea a la logverosimilitud definida de la siguiente manera

$$l(\beta_0, \beta^\top) = \sum_{i=1}^{n} y_i(\beta_0 + \beta^\top x_i) - \log[1 + \exp(\beta_0 + \beta^\top x_i)] \tag{2.24}$$

Obteniendo las siguientes ecuaciones

$$\begin{aligned} &\sum_{i=1}^{n} y_i - P(\mathtt{Y}=1|\mathbf{x}) = 0 \\ &\sum_{i=1}^{n} x_{ij}[y_i - P(\mathtt{Y}=1|\mathbf{x})] = 0, \ j = 1, \ldots, p. \end{aligned} \tag{2.25}$$

Siendo $(\widehat{\beta}_0, \ldots, \widehat{\beta}_p)$ el vector de parámetros que cumplen con las ecuaciones anteriores. La desviación estándar estimada de cada coeficiente es la siguiente

$$\widehat{\mathrm{SE}}(\widehat{\beta}_i) = \sqrt{\widehat{\mathrm{Var}}(\widehat{\beta}_i)}$$

Se ajustó un modelo de regresión logística al conjunto de datos `DatosCancer` utilizando al vector de variables explicativas $\mathbf{x}^\top = (\mathtt{x}_6, \mathtt{x}_7)$ y variable respuesta `Y`. La variable $\mathtt{x}_7$ cuenta con dos niveles. Los coeficientes estimados se encuentran en la Tabla 2.4 y su *logit* estimado es el siguiente

$$logit[\widehat{P}(\mathtt{Y} = 1|\mathbf{x})] = 1.1022 - 0.0009\mathtt{x}_6 - 0.7487(\mathtt{x}_7 = 1) \tag{2.26}$$

Tabla 2.4: Coeficientes estimados para el conjunto de datos `DatosCancer`. Se exponen los coeficientes para los casos $\mathtt{x}_7 = 0$ y $\mathtt{x}_7 = 1$, cuando $\mathtt{x}_7 = 0$ su coeficiente es 0.

	$\widehat{\beta}$	Desv. Est.	z *value*	Pr(>|z|)
Intercepto	1.1022	0.3677	2.99	0.0027
$\mathtt{x}_6$	-0.0009	0.0003	-3.03	0.0024
$\mathtt{x}_7 = 1$	-0.7487	0.3248	-2.30	0.0211

Como ya se dijo anteriormente, las pruebas sobre la significancia estadística del modelo así como sus intervalos de confianza no son utilizados para medir el poder predictivo de la regresión logística múltiple y únicamente están presentados en este trabajo como una extensión de conocimientos.

Pruebas sobre la significancia estadística del modelo

En este caso se lleva a cabo la prueba de la razón de verosimilitud de la misma manera que se expuso en la Subsección 2.3.1. La prueba recae sobre la estadística G (2.17) donde la prueba de hipótesis nula define que los p coeficientes son iguales a cero. En esta ocasión, G se distribuye ji-cuadrada con p grados de libertad.
Del modelo expuesto en la Tabla 2.4 se evaluará si al menos un coeficiente es diferente de cero. Primero se obtiene la logverosimilitud que tiene un valor de -121.024 y se calcula la logverosimilitud para el modelo constante, obteniendo un valor de -128.538. Una vez teniendo los valores se procede a calcular G

$$G = -2[-128.538 - (-121.024)] = 15.02$$

Obteniendo un *p-value* de $P[\chi^2(2) > 15.02] = 0.000548 \leq 0.05$. Concluyendo que, con una significancia de 0.05, se encontró evidencia estadística que al menos un coeficiente del modelo es diferente de cero.
El ejemplo anterior puede ser extrapolado a probar si un modelo es mejor que otro. Se calcularían las logverosimilitudes de ambos modelos para poder obtener a la estadística G y finalmente se obtendría el *p-value* de una distribución ji-cuadrada con w grados de libertad, donde w es la diferencia en el número de coeficientes entre los modelos.

Prueba Wald

Se puede realizar la prueba Wald a cada coeficiente estimado. En la Tabla 2.4 se encuentra esta prueba en la columna Pr(>|z|) que es el resultado del siguiente cociente

$$W_i = \frac{\widehat{\beta}_i}{\widehat{\mathrm{SE}}(\widehat{\beta}_i)}$$

Con $i = 1, \ldots, p$

Intervalos de Confianza

Los métodos de estimación de intervalos de confianza para los coeficientes, el *logit* y las probabilidades estimadas son los mismos que en la Subsección 2.3.1.
Suponga los coeficientes de la Tabla 2.4. El intervalo de 95 % de confianza para el coeficiente de la variable $\mathtt{x}_6$ es

$$(-0.0009 - 1.96 \times 0.0003, -0.0009 + 1.96 \times 0.0003) = (-0.0014, -0.0003)$$

Para calcular un intervalo de confianza del *logit* estimado definido por

$$logit[\widehat{P}(\mathtt{Y} = 1|\mathbf{x})] = \widehat{\beta}_0 + \widehat{\beta}_1\mathtt{x}_1 + \cdots + \widehat{\beta}_p\mathtt{x}_p \tag{2.27}$$

Se requiere al estimador de la varianza del *logit* estimado, la cual está definida como (Hosmer et al., 2013, p.43)

$$\widehat{\mathrm{Var}}\{logit[\widehat{P}(\mathtt{Y} = 1|\mathbf{x})]\} = \sum_{i=0}^{p} \mathtt{x}_i^2 \widehat{\mathrm{Var}}(\widehat{\beta}_i) + \sum_{i=0}^{p}\sum_{j=i+1}^{p} 2\mathtt{x}_i\mathtt{x}_j \widehat{\mathrm{Cov}}(\widehat{\beta}_i, \widehat{\beta}_j) \tag{2.28}$$

Con $\mathtt{x}_0 = 1$.
Los puntos extremos del intervalo de $100(1-\alpha)\,\%$ de confianza para el estimador del *logit* son los siguientes

$$logit[\widehat{P}(\mathtt{Y} = 1|\mathbf{x})] \pm z_{1-\alpha/2}\widehat{\mathrm{SE}}\{logit[\widehat{P}(\mathtt{Y} = 1|\mathbf{x})]\}$$

Donde $\widehat{\mathrm{SE}}\{logit[\widehat{P}(\mathtt{Y} = 1|\mathbf{x})]\}$ es la raíz positiva de la ecuación (2.28).

Considerando al modelo de la Tabla 2.4, el *logit* estimado para una observación $x = (22, 1)$ es

$$logit[\widehat{P}(\mathtt{Y} = 1|\mathbf{x})] = 1.1022 - 0.0009 \times 22 - 0.7487 \times 1 = 0.3337 \tag{2.29}$$

Con la matriz de covarianzas estimadas expuesta a continuación, se puede calcular la varianza estimada del *logit*.

$$\begin{array}{c} \\ \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{array} \begin{array}{c} \begin{array}{ccc} \widehat{\beta}_0 & \widehat{\beta}_1 & \widehat{\beta}_2 \end{array} \\ \begin{pmatrix} 0.13 & -8 \times 10^{-5} & -0.07 \\ -8 \times 10^{-5} & 9 \times 10^{-8} & 9 \times 10^{-6} \\ -0.07 & 9 \times 10^{-6} & 0.10 \end{pmatrix} \end{array} \tag{2.30}$$

Tabla 2.5: Coeficientes estimados para el conjunto de datos `DatosCancer`. Se incluye la interacción entre las variables $\mathtt{x}_2$ y $\mathtt{x}_3$ denotada por $\mathtt{x}_2{:}\mathtt{x}_3$.

	$\widehat{\beta}$	Desv. Est.	z *value*	Pr(>|z|)
Intercepto	1.9478	0.5742	3.39	0.0006
$\mathtt{x}_2$	-0.6719	0.2073	-3.24	0.0011
$\mathtt{x}_5$	-0.0014	0.0003	-4.11	3×10^{-5}
$\mathtt{x}_2{:}\mathtt{x}_3$	0.0370	0.0146	2.53	0.01125

La varianza estimada sería

$$\begin{aligned}\widehat{\mathrm{Var}}\{logit[\widehat{P}(\mathtt{Y}=1|\mathbf{x})]\} &= 0.13 + 22^2 \times (9\times 10^{-8}) + 1^2 \times 0.10 \\ &\quad + 2 \times 22 \times (-8 \times 10^{-5}) + 2 \times 1 \times (-0.07) \\ &\quad + 2 \times 22 \times 1 \times (9 \times 10^{-6}) = 0.08\end{aligned}$$

Finalmente, el intervalo de 95 % de confianza para el *logit* estimado es

$$(0.3337 - 1.96 \times 0.28, 0.3337 + 1.96 \times 0.28) = (-0.2115, 0.8825)$$

Se observa que el 0 está contenido en el intervalo, por lo tanto se puede suponer que $(\beta_1, \beta_2) = (0, 0)$. El ejemplo anterior tuvo como propósito ejemplificar la metodología para el cálculo de un intervalo de confianza para el *logit* estimado.

Clasificación

El método de clasificación de observaciones es análogo al expuesto en la Subsección 2.3.1. La Tabla 2.5 expone a los coeficientes estimados donde se logran apreciar al coeficiente relacionado con la interacción entre las variables $\mathtt{x}_2$ y $\mathtt{x}_3$. El *logit* estimado del modelo es

$$logit[\widehat{P}(\mathtt{Y}=1|\mathbf{x})] = 1.9478 - 0.6719\mathtt{x}_2 - 0.0014\mathtt{x}_5 + 0.0370\mathtt{x}_2\mathtt{x}_3 \tag{2.31}$$

Suponga una observación $x = (\mathtt{x}_2 = 2, \mathtt{x}_3 = 15, \mathtt{x}_5 = 22)$, el *logit* estimado sería

$$\begin{aligned}logit\{\widehat{P}[\mathtt{Y}=1|\mathbf{x}=(2,15,22)]\} &= 1.9478 - 0.6719 \times 2 - 0.0014 \times 22 \\ &\quad + 0.0370 \times 2 \times 15 = 1.68\end{aligned}$$

Con una probabilidad *a posteriori* estimada de

$$\widehat{P}(\mathtt{Y}=1|\mathbf{x}) = \frac{\exp(1.68)}{1+\exp(1.68)} \approx 0.84$$

Concluyendo que, a la observación $x = (2, 15, 22)$, se clasificará dentro de la clase 1.

Se explicó el método de clasificación, sin embargo, el poder predictivo de la regresión logística será abordado en el Capítulo 3.

2.4. Selección de Variables

2.4.1. *Stepwise Selection*

El método de *stepwise selection* es una alternativa al algoritmo de *best subset selection* que no se puede aplicar a un modelo con un número significativo de variables explicativas. Para p variables, *best subset selection* ajusta 2^p modelos y *stepwise selection* a lo más $1+p(p+1)/2$ modelos. Si $p = 20$ *best subset selection* ajusta $1,048,576$ modelos y *stepwise selection* 211 (James et al., 2013, p.229).

En este trabajo se abordará únicamente al algoritmo de *backward stepwise selection* aplicado a la regresión logística, el algoritmo de *forward stepwise selection* así como el de *best subset selection* aplicados a otros métodos, se pueden consultar en *An Introduction to Statistical Learning: with Applications in R* de James, Witten, Hastie & Tibshiani, capítulo 6.
Antes de comenzar con el algoritmo, se expresarán a los siguientes criterios

- *Akaike Information Criterion* (AIC)

  La expresión del criterio AIC es el siguiente (Venables & Ripley, 2002, p.174)

  $$\text{AIC} = -2l(\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p) + 2p$$

  Donde p es el número de variables explicativas presentes en el modelo y $l(\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p)$ la logverosimilitud.

- *Bayesian Information Criterion* (BIC)

  La expresión del criterio BIC es el siguiente (Hastie et al., 2009, p.233)

  $$\text{BIC} = -2l(\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p) + p\log(n)$$

  Donde n es el número de observaciones, p el número de variables explicativas presentes en el modelo y $l(\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p)$ la logverosimilitud.

Backward Stepwise Selection

Suponga que se tienen p variables explicativas con una variable respuesta. El algoritmo *backward stepwise selection* comienza con un modelo de p variables explicativas y luego de manera iterativa va quitando una variable a la vez. La variable que se va quitando es la menos útil en términos de una métrica, por ejemplo: AIC, BIC o en el cálculo del poder predictivo. Se necesita que el número de observaciones sea mayor al número de variables explicativas y el algoritmo se expone a continuación (James et al., 2013, p.231).
Para el caso de la regresión logística se utilizan los criterios AIC y BIC.

Considere al conjunto de datos `DatosCancer`. Se empleará una selección de variables utilizando *backward stepwise selection*. En la Tabla 2.6 se puede observar que la prueba Wald arroja que todos los coeficientes son significativos para la variable respuesta. Al comparar el modelo de la Tabla 2.6 con el de la Tabla 2.8, se destaca

Algoritmo empleado en *backward stepwise selection*

1. Sea M_p un modelo completo con p variables explicativas.
2. Para $k = p, p-1, \ldots, 1$:
   a) Defina a M_k como el modelo con k variables explicativas. Considere a los k modelos que contienen a todas las variables explicativas en M_k menos a uno, para un total de $k-1$ variables explicativas.
   b) Escoja el mejor modelo dentro de los k modelos y defínalo como M_{k-1}. Para la regresión lineal, el criterio de mejor modelo es el que tenga el mayor valor de R^2 pero puede ser algún otro criterio.

el resultado de la prueba Wald para los coeficientes estimados. Esto repercute al predecir a la variable respuesta `Y`. Las diferencias de los errores de predicción se pueden observar en la Tabla 2.9. El error de predicción mejoró al eliminar a las variables: $\mathtt{x}_1, \mathtt{x}_4, \mathtt{x}_6, \mathtt{x}_7, \mathtt{x}_8$ y $\mathtt{x}_9$. Al aplicar el algoritmo *backward stepwise selection* con el criterio BIC, ver Tabla 2.7, se obtuvo un modelo con una sola variable explicativa y un error de predicción de 39.24 %. El modelo con criterio AIC, ver Tabla 2.6, obtuvo un error de predicción de 28.49 %. Este es un ejemplo donde el mejor modelo no siempre es el que contiene menos coeficientes.

Tabla 2.6: Coeficientes estimados para el modelo de *backward stepwise selection* utilizando el criterio AIC. Se observa que todos los coeficientes son estadísticamente significativos según la prueba Wald.

	$\widehat{\beta}$	Desv. Est.	z *value*	Pr(>|z|)
Intercepto	1.0376	0.6295	1.65	0.0993
$\mathtt{x}_2$	-0.3090	0.1491	-2.07	0.0383
$\mathtt{x}_3$	0.0848	0.0438	1.94	0.0529
$\mathtt{x}_5$	-0.0013	0.0003	-3.89	0.0001

Tabla 2.7: Coeficientes estimados para el modelo de *backward stepwise selection* utilizando el criterio BIC.

	$\widehat{\beta}$	Desv. Est.	z *value*	Pr(>|z|)
Intercepto	0.6329	0.2557	2.47	0.0133
$\mathtt{x}_5$	-0.0009	0.0002	-3.47	0.0005

Tabla 2.8: Coeficientes estimados para el modelo aditivo (2.35) del conjunto de datos `DatosCancer`. Se observa que no todos los coeficientes son estadísticamente significativos según la prueba Wald.

	$\widehat{\beta}$	Desv. Est.	z *value*	Pr(>|z|)
Intercepto	1.7713	1.2522	1.41	0.1572
$\mathtt{x}_1$	-0.0110	0.0193	-0.57	0.5671
$\mathtt{x}_2$	-0.2605	0.1845	-1.41	0.1579
$\mathtt{x}_3$	0.0836	0.0447	1.87	0.0613
$\mathtt{x}_4$	0.0001	0.0007	0.15	0.8833
$\mathtt{x}_5$	-0.0012	0.0005	-2.56	0.0105
$\mathtt{x}_6$	-0.0003	0.0005	-0.67	0.5038
$\mathtt{x}_7$=1	-0.3677	0.4209	-0.87	0.3824
$\mathtt{x}_8$=1	-0.2853	0.3273	-0.87	0.3833
$\mathtt{x}_9$=1	0.3108	0.3413	0.91	0.3625

Tabla 2.9: Comparación de los errores aparentes obtenidos (%) entre los modelos expuestos en la Tabla 2.6 y 2.8.

	Error Aparente		
Modelo de regresión logística	Global	Clase 0	Clase 1
Aditivo	32.25	32.32	32.18
Stepwise selection AIC	28.49	25.25	32.18

2.5. Métodos de Regularización

Los métodos de Selección de Variables son procesos discretos, es decir, las variables explicativas son consideradas o eliminadas de lo modelos. Como consecuencia se presenta una alta varianza en los resultados. Los métodos de regularización no sufren de esta variabilidad (Hastie et al., 2009, p.61).

Los métodos que se presentan a continuación pueden ser aplicados a modelos como la regresión lineal, análisis discriminante, entre otros. En este trabajo únicamente se abordará el caso de la regresión logística, si se busca conocer más acerca de la aplicación en otros modelos, se le recomienda al lector *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* sección 3.4 o *Statistical Learning with Sparsity: The Lasso and Generalizations*.

En la regresión lineal, se estiman a los parámetros del modelo utilizando la función de *mínimos cuadrados*, estos parámetros normalmente serán diferentes de cero dificultando la interpretación del modelo si p es grande. Si $n > p$, habrá una infinidad de parámetros estimados que hagan cero a la función de mínimos cuadrados ocasionando *overfitting*, para evitarlo, se regulariza imponiendo una penalización en el tamaño de los parámetros estimados (Hastie et al., 2020, p.2).

En la regresión logística, esta regularización se aplica a la función de logverosimilitud (2.24) creando el siguiente problema de optimización

$$\begin{gathered}\underset{(\beta_0,\beta^\top)}{\text{argmín}} \left\{ \frac{1}{n}\sum_{i=1}^{n} \left\{ y_i(\beta_0+\beta^\top x_i) - \log[1+\exp(\beta_0+\beta^\top x_i)] \right\} \right\} \\ \text{restringido a } \sum_{j=1}^{p} |\beta_j|^q \leq t \end{gathered} \tag{2.32}$$

Con $q \geq 0$. Aunque q puede ser cualquier valor, si $q < 1$ el problema no es convexo y hace que la minimización sea difícil computacionalmente. El valor $q = 1$ es el valor más pequeño que hace convexo al problema (Hastie et al., 2020, p.2).
El problema anterior también puede ser expresado de la siguiente manera (Hastie et al., 2020, p.9)

$$\widehat{\beta} = \underset{(\beta_0,\beta^\top)}{\text{argmín}} \left\{ \frac{1}{n}\sum_{i=1}^{n} \left\{ y_i(\beta_0+\beta^\top x_i) - \log[1+\exp(\beta_0+\beta^\top x_i)] \right\} + \lambda \sum_{j=1}^{p} |\beta_j|^q \right\} \tag{2.33}$$

Donde $x_i = (x_{i1}, \ldots, x_{ip})$, n el número de observaciones, p el número de variables presentes en el modelo, $\lambda \geq 0$ un parámetro a determinar y $q \geq 0$. Si $q = 2$, corresponde a *ridge regression* y si $q = 1$ a *lasso*.

2.5.1. *Ridge Regression*

Para el caso de la regresión logística, este método estima a los parámetros β_0 y $\beta^\top = (\beta_1, \ldots, \beta_p)$ del problema de optimización (James et al., 2013, p.243)

$$\widehat{\beta}^R = \underset{(\beta_0,\beta^\top)}{\text{argmín}} \left\{ \frac{1}{n}\sum_{i=1}^{n} \left\{ y_i(\beta_0+\beta^\top x_i) - \log[1+\exp(\beta_0+\beta^\top x_i)] \right\} + \lambda \sum_{j=1}^{p} |\beta_j|^2 \right\} \tag{2.34}$$

Cuando se minimiza a la ecuación (2.34) también se minimiza a $\lambda \sum \beta_j^2$ conocida como la *penalización de regularización*, alcanzando su mínimo cuando los valores de $\beta_1, \ldots, \beta_p$ se encojen.

Ridge regression genera vectores de coeficientes $\widehat{\beta}^R$ dependiendo del valor de λ que es determinado utilizando *cross-validation*. Se escoge al valor de λ que haya arrojado el error de clasificación más pequeño. Finalmente, se vuelve a estimar el modelo utilizando el valor de λ óptimo.

En la Figura 2.3 se observa el resultado de aplicar *ridge regression* al modelo aditivo del conjunto de datos `DatosCancer` expresado a continuación utilizando la notación de Wilkinson

$$\mathtt{Y} \sim \mathtt{x}_1 + \mathtt{x}_2 + \mathtt{x}_3 + \mathtt{x}_4 + \mathtt{x}_5 + \mathtt{x}_6 + \mathtt{x}_7 + \mathtt{x}_8 + \mathtt{x}_9 \tag{2.35}$$

En la gráfica de la derecha se observan los coeficientes estimados, notando que en el extremo izquierdo de la gráfica el valor de λ es cercano a cero, es decir, no hay

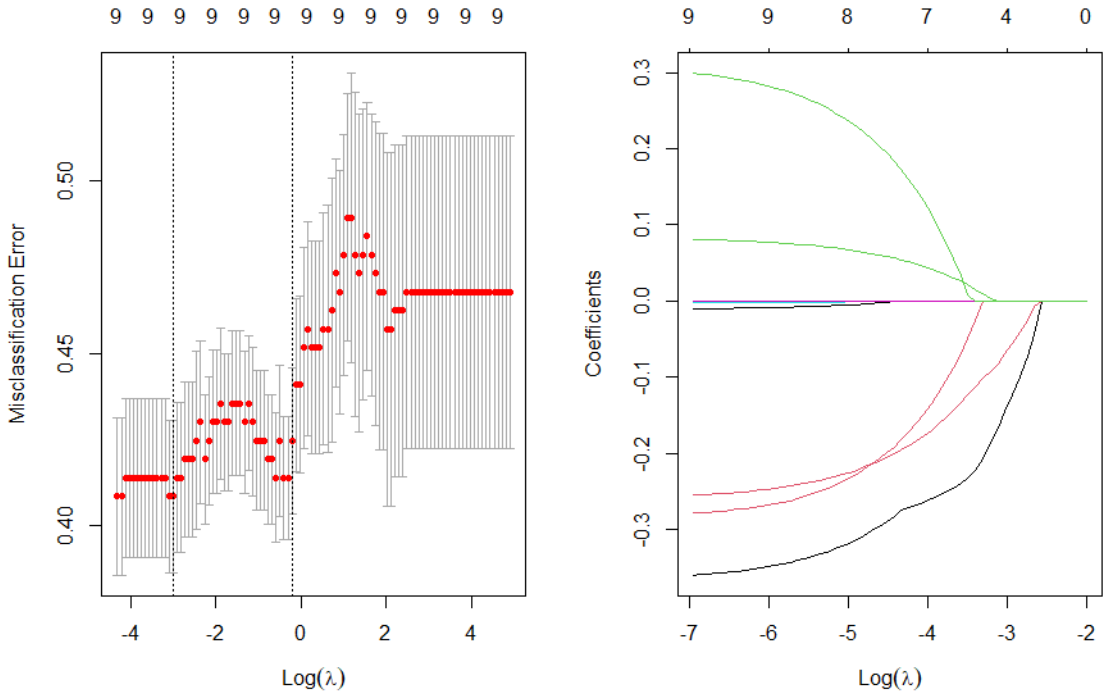

Figura 2.3: Resultado del método *ridge regression*. Del lado izquierdo se observan los errores obtenidos por *10-fold cross-validation* con respecto a los valores de λ. Del lado derecho se observan los valores de los coeficientes estimados $\widehat{\beta}^R$ con respecto al valor de λ.

regularización y por ende los coeficientes son los mismos que en la Tabla 2.8.
En la gráfica de la izquierda se expone la metodología *cross-validation* para encontrar el valor óptimo de λ, en este caso se utilizó *10-fold cross-validation*. Los coeficientes estimados $\widehat{\beta}^R$ se exponen en la Tabla 2.10 y se obtuvo una tasa de error aparente de 34.41 %.

Tabla 2.10: Coeficientes estimados $\widehat{\beta}^R$ al aplicar *ridge regression* al modelo aditivo (2.35) del conjunto de datos `DatosCancer`.

	$\widehat{\beta}$		$\widehat{\beta}$
Intercepto	1.3735	x_5	-0.0006
x_1	-0.0047	x_6	-0.0004
x_2	-0.2008	$\mathrm{x}_7 = 1$	-0.3428
x_3	0.0488	$\mathrm{x}_8 = 1$	-0.2510
x_4	-6×10^{-5}	$\mathrm{x}_9 = 1$	0.2416

2.5.2. *Lasso*

Lasso es un método que estima a los coeficientes, para el caso de regresión logística, del siguiente problema de optimización (James et al., 2013, p.243)

$$\widehat{\beta}^{L} = \underset{(\beta_0,\beta^\top)}{\text{argmín}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left\{ y_i(\beta_0 + \beta^\top x_i) - \log[1 + \exp(\beta_0 + \beta^\top x_i)] \right\} + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{2.36}$$

Se tiene el objetivo de determinar el valor óptimo para λ utilizando *cross-validation*. En la gráfica del lado izquierdo de la Figura 2.4 se puede observar los errores obtenidos con respecto al valor de λ utilizando *10-fold cross-validation*, la gráfica del lado derecho expone el comportamiento de los coeficientes con respecto al valor de λ. En la Tabla 2.11 se presentan los coeficientes estimados al aplicar *lasso*, en este caso el coeficiente de la variable $\mathtt{x}_4$ resultó ser cero.

Una diferencia entre los métodos regularizados y los no regularizados, es que no existe una forma de evaluar la significancia estadística de los coeficientes en los métodos regularizados. La Tabla 2.5 expone la desviación estándar estimada de los coeficientes estimados pero la Tabla 2.11 no lo hace, se tiene que utilizar algún método de remuestreo como *bootstrap* para calcular la desviación estándar estimada (Hastie et al., 2020, p.12).

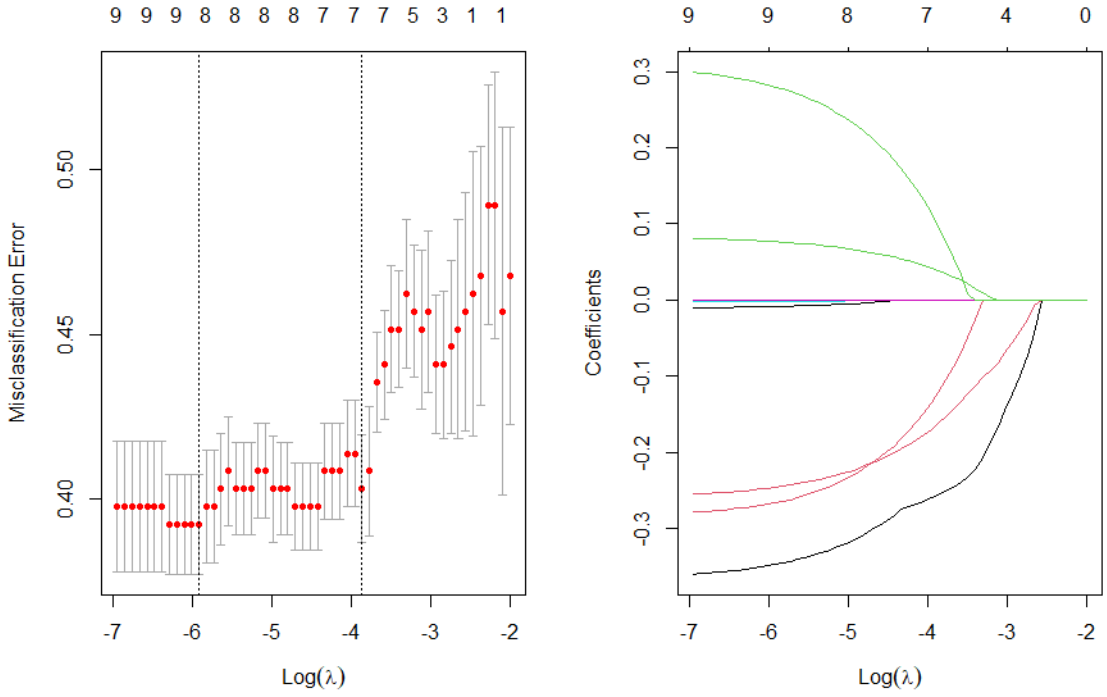

Figura 2.4: Resultado del método *lasso*. La gráfica del lado izquierdo expone los errores obtenidos por *10-fold cross-validation* con respecto a los valores de λ, la del lado derecho, los valores de los coeficientes estimados $\widehat{\beta}^{R}$ con respecto al valor de λ.

Tabla 2.11: Coeficientes estimados $\widehat{\beta}^L$ al aplicar *lasso* al modelo aditivo (2.35) del conjunto de datos `DatosCancer`.

	$\widehat{\beta}$		$\widehat{\beta}$
Intercepto	1.6119	$\mathtt{x}_5$	-0.0011
$\mathtt{x}_1$	-0.0084	$\mathtt{x}_6$	-0.0002
$\mathtt{x}_2$	-0.2456	$\mathtt{x}_7 = 1$	-0.3470
$\mathtt{x}_3$	0.0766	$\mathtt{x}_8 = 1$	-0.2665
$\mathtt{x}_4$	0.0000	$\mathtt{x}_9 = 1$	0.2797

La gran diferencia entre *lasso* y *ridge regression*, es que *lasso* puede hacer cero a los coeficientes. Esta característica es consecuencia de la norma de la penalización de regularización en *lasso* que es l_1.
La Figura 2.5 fue tomada de *Statistical Learning with Sparsity: The Lasso and Generalizations* página 11 y expone las diferencias entre los métodos, el área azul corresponde a la región delimitada por la restricción expuesta en (2.32), es decir, $||\beta||_1 \leq t$ para *lasso* (gráfica izquierda) y $||\beta||_2 \leq t$ para *ridge regression* (gráfica derecha). Las elipses en color rojo corresponden a la función a la que se le aplica la regularización con valores constantes y $\widehat{\beta}$ a la solución que minimiza a la función. La Figura 2.5 expone el caso para la regresión lineal con $p = 2$, las elipses corresponden a la función *Residual Sum of Squares* (RSS) con valores constantes y entre más se alejan de $\widehat{\beta}$ el valor de RSS aumenta. Las soluciones para *lasso* y *ridge regression* son los primeros puntos donde se intersecan las elipses y las áreas azules.
El área de restricción de *lasso* es un rombo con esquinas y, si la solución ocurre en una de ellas, se produce un parámetro igual a cero. Cuando $p > 2$ el rombo se convierte en un romboide que incrementa la posibilidad que los parámetros estimados sean cero, a comparación de *ridge regression* que su área de restricción es un círculo sin esquinas (Hastie et al., 2020, p.12).

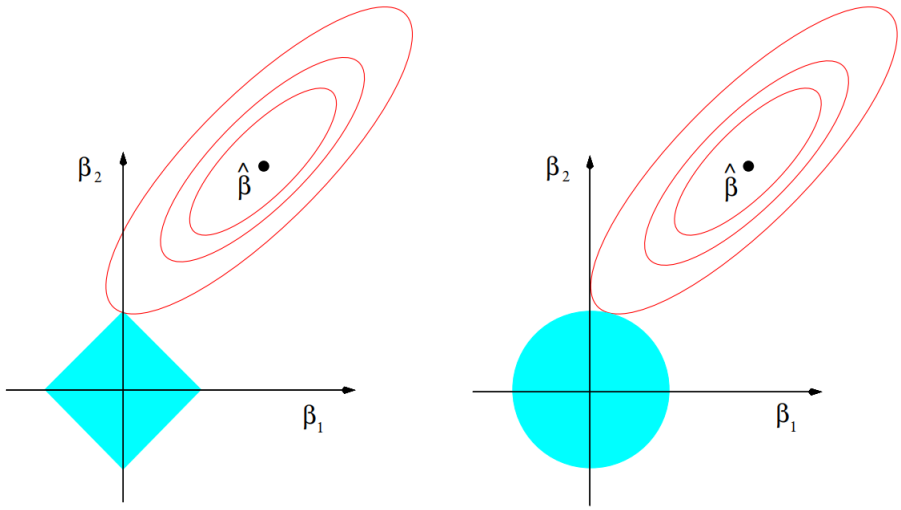

Figura 2.5: Diferencia entre las soluciones de *lasso* y *ridge regression* para $p = 2$.
Fuente: *Statistical Learning with Sparsity: The Lasso and Generalizations* p.11.

2.5.3. *Elastic Net*

El método de *Elastic Net* se le atribuye a Zou & Hastie (2005) y es una combinación entre las regularizaciones *lasso* y *ridge regression*. Este método se puede aplicar a diferentes modelos, en este caso se abordará únicamente regresión logística. La regularización *elastic net* es equivalente al siguiente problema de optimización (Zou & Hastie, 2005)

$$\widehat{\beta}^{EN} = \underset{(\beta_0,\beta^\top)}{\text{argmáx}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left\{ y_i(\beta_0 + \beta^\top x_i) - \log[1 + \exp(\beta_0 + \beta^\top x_i)] \right\} - \lambda P_\alpha(\beta) \right\} \tag{2.37}$$

Donde $P_\alpha(\beta)$ es la penalización de *elastic net* dada por la siguiente expresión (Zou & Hastie, 2005)

$$P_\alpha(\beta) = (1-\alpha)||\beta||_1 + \alpha||\beta||_2^2 \tag{2.38}$$

Con $\alpha \in [0, 1]$.
Notando la combinación entre las penalizaciones *lasso* y *ridge regression*. Cuando $\alpha = 1$, corresponde a *ridge regression* y cuando $\alpha = 0$ a *lasso*.

En textos más recientes como *Computer Age Statistical Inference, Student Edition: Algorithms, Evidence, and Data Science*, se expresa a la penalización como (Efron & Hastie, 2021, p.316)

$$P_\alpha(\beta) = \frac{1}{2}(1-\alpha)||\beta||_2^2 + \alpha||\beta||_1 \tag{2.39}$$

Donde el factor $1/2$ es utilizado por conveniencia matemática.
En esta expresión, si $\alpha = 1$ se obtiene la penalización *lasso* y si $\alpha = 0$ la penalización *ridge regression*. En este trabajo se utilizará la expresión (2.39).

Cuando $\alpha < 1$ y $\lambda > 0$, existe una única solución sin importar la correlación entre las variables explicativas. El parámetro α debe de ser determinado y la forma de hacerlo es que el usuario simplemente lo seleccione o usando una malla de valores de α aplicando *cross-validation* (Hastie et al., 2020, p.57).
En la Figura 2.6 se observa el comportamiento de los coeficientes con respecto al valor de λ para *lasso, ridge regression* y *elastic net*.

La Figura 2.7 fue tomada de *Statistical Learning with Sparsity: The Lasso and Generalizations* página 58 y compara las regiones de restricción para *elastic net* (gráfica izquierda) y *lasso* (gráfica derecha) con $p = 3$. Se observa que la región de *elastic net* comparte características de *lasso* y *ridge regression*: las esquinas y bordes afilados que fomentan la selección de variables, es decir, hacer algunos coeficientes cero y, los contornos curvos, que fomentan que variables altamente correlacionadas compartan coeficientes (Hastie et al., 2020, p.57).

En la Tabla 2.12 se exponen a los coeficientes estimados al aplicar *elastic net*, con $\alpha = 0.5$, al modelo aditivo (2.35) del conjunto de datos `DatosCancer` y en la Tabla 2.13, los errores aparentes obtenidos por los modelos de selección de variables y regularización. Para los métodos de regularización se seleccionó a λ utilizando *10-fold cross-validation* basado en el porcentaje de clasificación errónea. Se seleccionó a la λ que arrojó el error de clasificación más bajo.

Tabla 2.12: Coeficientes estimados al aplicar *elastic net*, con $\alpha = 0.5$, al modelo aditivo (2.35) del conjunto de datos `DatosCancer`.

	$\widehat{\beta}$		$\widehat{\beta}$
Intercepto	1.6150	$\mathtt{x}_5$	-0.0010
$\mathtt{x}_1$	-0.0084	$\mathtt{x}_6$	-0.0003
$\mathtt{x}_2$	-0.2444	$\mathtt{x}_7 = 1$	-0.3499
$\mathtt{x}_3$	0.0751	$\mathtt{x}_8 = 1$	-0.2687
$\mathtt{x}_4$	5e-6	$\mathtt{x}_9 = 1$	0.2816

Tabla 2.13: Errores aparentes obtenidos (%) al aplicar *stepwise selection* y los métodos de regularización al modelo aditivo (2.35) del conjunto de datos `DatosCancer`.

	Error Aparente		
Modelo de regresión logística	Global	Clase 0	Clase 1
Aditivo	32.25	32.32	32.18
Stepwise selection AIC	28.49	25.25	32.18
Stepwise selection BIC	39.24	43.43	34.48
Ridge regression	34.41	28.28	41.37
Lasso	32.79	31.31	34.48
Elastic net	31.72	31.31	32.18

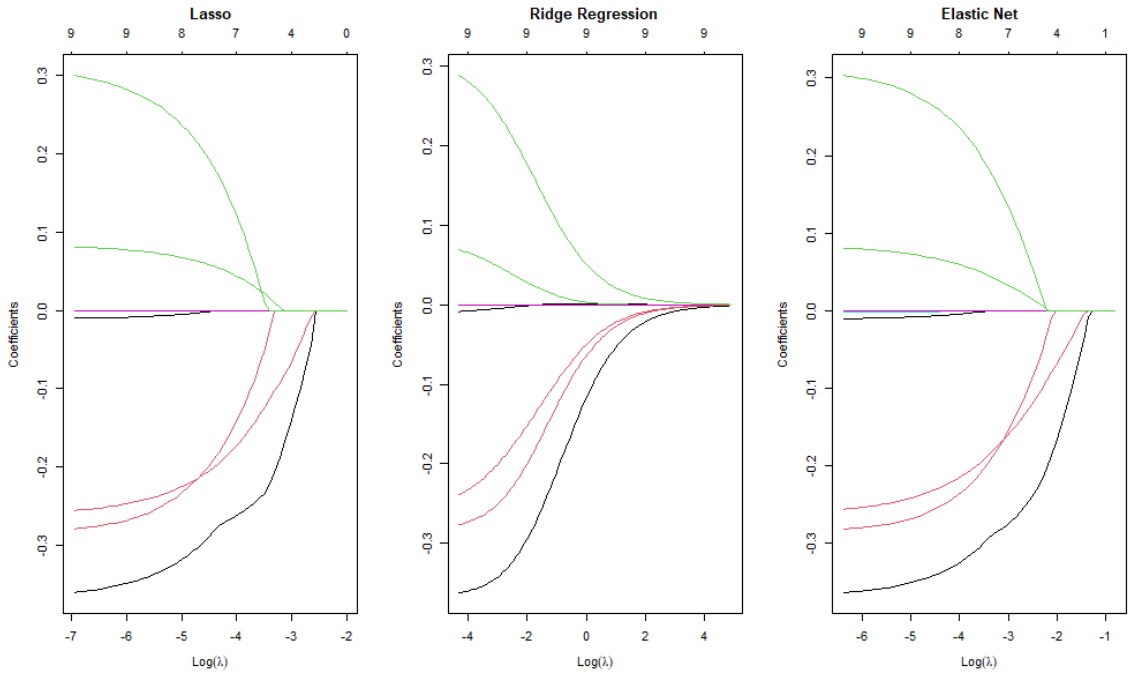

Figura 2.6: Comparación entre los coeficientes estimados y el valor de λ en los métodos de regularización *lasso* ($\alpha = 1$), *ridge regression* ($\alpha = 0$) y *elastic net* ($\alpha = 0.3$), respectivamente.

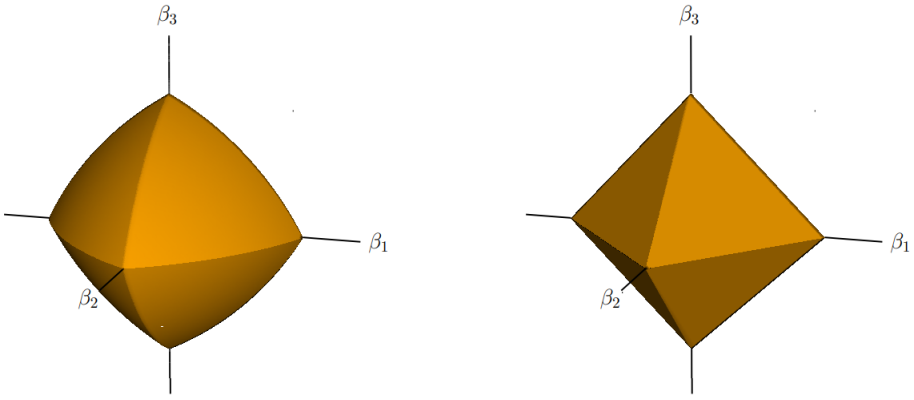

Figura 2.7: Regiones de restricción para *elastic net* con $\alpha = 0.7$ (gráfica izquierda) y *lasso* (gráfica derecha) con $p = 3$.
Fuente: *Statistical Learning with Sparsity: The Lasso and Generalizations* p.58.

2.6. Análisis Discriminante

2.6.1. Análisis Discriminante de Fisher

El Análisis Discriminante, en su forma lineal para dos clases y sin supuestos distribucionales, se le atribuye a Fisher (1936).
La función Discriminante Lineal de Fisher consta de la combinación lineal de las p variables explicativas que mejor separen a las medias de dos conjuntos en un espacio p-dimensional. Se denota de la siguiente manera

$$LDF = \alpha^\top \mathbf{x} = \alpha_1 \mathbf{x}_1 + \cdots + \alpha_p \mathbf{x}_p \tag{2.40}$$

Si $K = k$, se estiman a los coeficientes del vector $\alpha^\top = (\alpha_1, \ldots, \alpha_p)$ de tal forma que

$$\underset{\alpha \in \mathbb{R}^p}{\text{argmáx}} \left\{ \frac{\alpha^\top B \alpha}{\alpha^\top W \alpha} \right\} \tag{2.41}$$

Con B la matriz de covarianzas entre clases y W la matriz de covarianzas dentro de las clases, ambas se expresan en (2.42), y n_k el número de observaciones en la k-ésima clase.

$$\begin{aligned} B =& \frac{1}{K-1} \sum_{k=1}^{K} n_k (\bar{x}_k - \bar{x})(\bar{x}_k - \bar{x})^\top \\ W =& \frac{1}{n-K} \sum_{k=1}^{K} \sum_{j=1}^{n_k} (x_{kj} - \bar{x}_k)(x_{kj} - \bar{x}_k)^\top \end{aligned} \tag{2.42}$$

Para maximizar la ecuación (2.41), se deriva con respecto a α y se iguala a cero

$$\begin{aligned} & \frac{\partial J(\alpha)}{\partial \alpha} = 0 \\ \implies & \frac{\partial}{\partial \alpha} \left(\frac{\alpha^\top B \alpha}{\alpha^\top W \alpha} \right) = 0 \\ \implies & (\alpha^\top W \alpha) \frac{\partial}{\partial \alpha} (\alpha^\top B \alpha) - (\alpha^\top B \alpha) \frac{\partial}{\partial \alpha} (\alpha^\top W \alpha) = 0 \\ \implies & (\alpha^\top W \alpha) 2B\alpha - (\alpha^\top B \alpha) 2W\alpha = 0 \\ \implies & B\alpha - \frac{\alpha^\top B \alpha}{\alpha^\top W \alpha} W\alpha = 0 \\ \implies & B\alpha - J(\alpha) W\alpha = 0 \\ \implies & B\alpha - \lambda W\alpha = 0 \end{aligned} \tag{2.43}$$

Se reduce a resolver la ecuación

$$(B - \lambda W)\alpha = 0 \tag{2.44}$$

Dada una nueva observación x, se le clasifica dentro una de las K clases con base a su puntaje discriminante $\alpha^\top x$, en particular, se le clasifica a x dentro de la clase k si:

$$|\alpha^\top x - \alpha^\top \bar{\mathbf{x}}_k| < |\alpha^\top x - \alpha^\top \bar{\mathbf{x}}_l|, \text{ para toda } l \neq k$$

El análisis discriminante de Fisher no supone ninguna distribución de los datos. Los siguiente métodos sí lo hacen, específicamente se abordará el caso cuando se supone a $f_k(x)$ como la función de densidad de $\mathbf{x}$, en la clase k, de una distribución Gaussiana Multivariada definida de la siguiente manera

$$f_k(x) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_k)^\top \Sigma_k^{-1}(x-\mu_k)\right] \tag{2.45}$$

Recordando el Teorema de Bayes, se concluye que

$$P(\mathtt{Y}=k|\mathbf{x}=x) = \frac{\pi_k f_k(x)}{\sum_{i=1}^K \pi_i f_i(x)} \propto \pi_k f_k(x) \tag{2.46}$$

2.6.2. Análisis Discriminante Lineal Gaussiano

El *Análisis Discriminante Lineal* (LDA) surge al suponer que todas las clases tienen una matriz de covarianzas igual, es decir, $\Sigma_k = \Sigma$ para toda k.
La regla de clasificación de Bayes clasifica a una observación x en la clase k que maximice la probabilidad *a posteriori*. En este caso se tiene lo siguiente

$$\underset{k}{\text{argmáx}}\{P(\mathtt{Y}=k|\mathbf{x}=x)\} = \underset{k}{\text{argmáx}}\{\pi_k f_k(x)\} \tag{2.47}$$

Aplicando la función logaritmo $\log[\pi_k f_k(x)]$ se obtiene lo siguiente

$$\underset{k}{\text{argmáx}}\{-\log[(2\pi)^{p/2}|\Sigma|^{1/2}] - \frac{1}{2}(x-\mu_k)^\top\Sigma^{-1}(x-\mu_k) + \log(\pi_k)\} \tag{2.48}$$

Ya que se busca maximizar sobre k, el término $-\log[(2\pi)^{p/2}|\Sigma|^{1/2}]$ es constante y puede ser descartado, obteniendo

$$\underset{k}{\text{argmáx}}\{x^\top\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^\top\Sigma^{-1}\mu_k - \frac{1}{2}x^\top\Sigma^{-1}x + \log(\pi_k)\} \tag{2.49}$$

De igual forma, el término $-\frac{1}{2}x^\top\Sigma^{-1}x$ es constante y se descarta. Finalmente se tiene lo siguiente

$$\underset{k}{\text{argmáx}}\{x^\top\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^\top\Sigma^{-1}\mu_k + \log(\pi_k)\} \tag{2.50}$$

En (2.50) se define a $x^\top\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^\top\Sigma^{-1}\mu_k + \log(\pi_k)$ como la función de *Discriminante Lineal* (2.51) para la k-ésima clase (Hastie et al., 2009, p.109)

$$\delta_k^L(x) = x^\top\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^\top\Sigma^{-1}\mu_k + \log(\pi_k) \tag{2.51}$$

Con frontera de decisión, entre las clases a y b, definida por $\{x : \delta_a^L(x) = \delta_b^L(x)\}$.
Utilizando los datos de entrenamiento se estiman a los siguientes parámetros

$$\begin{aligned}
\widehat{\pi}_k &= \frac{n_k}{n} \\
\widehat{\mu}_k &= \frac{1}{n_k}\sum_{i:y_i=k} x_i \\
\widehat{\Sigma} &= \frac{1}{n-K}\sum_{k=1}^K \sum_{i:y_i=k} (x_i-\widehat{\mu}_k)(x_i-\widehat{\mu}_k)^\top
\end{aligned}$$

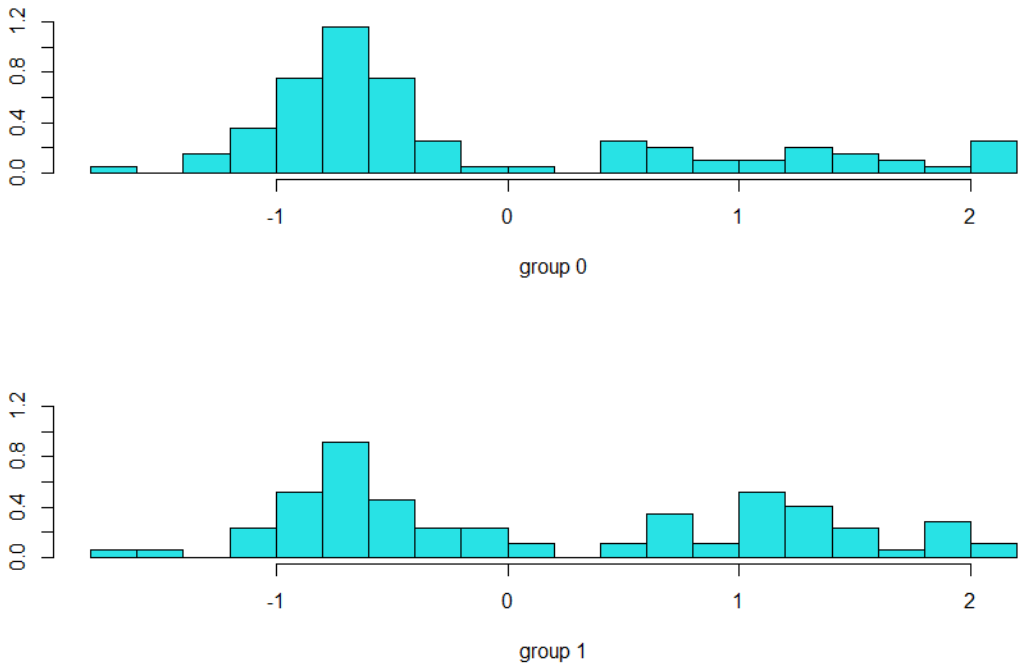

Figura 2.8: Histogramas de los valores obtenidos al aplicar (2.52) al conjunto de datos `DatosCancer`. Los histogramas se empalman, por consecuencia, no se puede distinguir ambas clases, obteniendo un error aparente de 41.39 %.

Clasificación

Al contar con K clases, se clasifica a una observación x a la clase k que maximice (2.51), es decir

$$\underset{k}{\text{argmáx}}\{\delta_k^L(x)\}$$

Si $K = 2$, se clasifica a la observación x a la clase 1 si $\delta_1(x) > \delta_0(x)$ y clase 0 en otro caso, o bien si

$$x^\top \widehat{\Sigma}^{-1}(\widehat{\mu}_1 - \widehat{\mu}_0) > \frac{1}{2}\widehat{\mu}_1^\top \Sigma^{-1}\widehat{\mu}_1 - \frac{1}{2}\widehat{\mu}_0^\top \widehat{\Sigma}^{-1}\widehat{\mu}_0 + \log(\widehat{\pi}_0) - \log(\widehat{\pi}_1)$$

Al aplicar LDA al conjunto de datos `DatosCancer`, utilizando a las variables explicativas $\mathtt{x}_1, \mathtt{x}_2$ y $\mathtt{x}_7$, se obtuvo una función discriminante lineal de Fisher

$$LDF = 0.010\mathtt{x}_1 - 0.439\mathtt{x}_2 - 1.320(\mathtt{x}_7 = 1) \tag{2.52}$$

En la Figura 2.8 se puede observar la interacción entre la ecuación (2.52) y el conjunto de datos. Se busca separar a las observaciones para posteriormente hacer una clasificación. En este caso, se obtuvo un error aparente de clasificación de 41.39 %.

2.6.3. Análisis Discriminante Cuadrático Gaussiano

En LDA se supuso que $\Sigma_k = \Sigma$ para toda k, al suponer que no son iguales, la ecuación (2.48) se define de la siguiente manera

$$\underset{k}{\text{argmáx}}\{-\log[(2\pi)^{p/2}|\Sigma_k|^{1/2}] - \frac{1}{2}(x-\mu_k)^\top \Sigma_k^{-1}(x-\mu_k) + \log(\pi_k)\} \tag{2.53}$$

Donde $-\log[(2\pi)^{p/2}]$ es constante y se puede descartar.
Se define a la función de *Discriminante Cuadrático* (QDA), para la k-ésima clase, de la siguiente forma (Hastie et al., 2009, p.110)

$$\delta_k^Q(x) = -\frac{1}{2}\log(|\Sigma_k|) - \frac{1}{2}(x-\mu_k)^\top \Sigma_k^{-1}(x-\mu_k) + \log(\pi_k) \tag{2.54}$$

Con frontera de decisión, entre las clases a y b, definida por $\{x : \delta_a^Q(x) = \delta_b^Q(x)\}$ y los parámetros estimados son

$$\begin{aligned}
\widehat{\pi}_k =& \frac{n_k}{n} \\
\widehat{\mu}_k =& \frac{1}{n_k}\sum_{i:y_i=k} x_i \\
\widehat{\Sigma}_k =& \frac{1}{n_k - 1}\sum_{i:y_i=k}(x_i - \widehat{\mu}_k)(x_i - \widehat{\mu}_k)^\top
\end{aligned}$$

Clasificación

Se clasifica a la observación x en la clase k que maximice (2.54), es decir

$$\underset{k}{\text{argmáx}}\{\delta_k^Q(x)\}$$

En general, LDA es útil cuando se tienen pocas observaciones y se requiere disminuir la varianza, en cambio, QDA es una mejor opción cuando se cuenta con una cantidad considerable de datos o cuando el supuesto que las clases comparten la matriz de covarianzas no es factible (James et al., 2013, p.153).

2.7. Métodos basados en árboles

2.7.1. Árboles de Decisión

La ventaja principal de los *Árboles de Decisión* o *Classification and Regression Trees* (CART) es su fácil interpretabilidad. En este trabajo se abordará el caso cuando la variable respuesta es binaria, para el caso cuando la variable respuesta sea continua se le recomienda al lector *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* de Hastie et al. (2009) subsección 9.2.2.

Árboles de Clasificación

Suponga que se tienen n observaciones independientes (x_i, y_i), p variables explicativas $\mathbf{x}^\top = (\mathtt{x}_1, \ldots \mathtt{x}_p)$ y la variable respuesta categórica $\mathtt{Y}$ con K clases. Los *Árboles de Clasificación* constan de lo siguiente

1. Particionar el espacio generado por las variables explicativas $\mathtt{x}_1, \ldots \mathtt{x}_p$ en J regiones disjuntas $R_1, \ldots, R_J$.
2. Una vez particionado el espacio, la clase estimada para la variable respuesta $\mathtt{Y}$, en la i-ésima observación, es la clase que más se repita dentro de la región donde se encuentre la observación.

Dado un nodo m que representa a la región R_m con n_m observaciones, la expresión de la proporción de las observaciones de la clase k en el nodo m es (Hastie et al., 2009, p.309)

$$\widehat{p}_{mk} = \frac{1}{n_m} \sum_{x_i \in R_m} I(y_i = k) \tag{2.55}$$

Con ayuda de (2.55) se clasifican a las observaciones x_i, del nodo m, a la clase que más proporción tenga dentro de la región R_m, es decir

$$k(m) = \underset{k}{\text{argmáx}}\{\widehat{p}_{mk}\} \tag{2.56}$$

Se define al error de clasificación como

$$\frac{1}{n_m} \sum_{x_i \in R_m} I[y_i \neq k(m)] = 1 - \widehat{p}_{mk(m)} \tag{2.57}$$

Esta expresión es la más utilizada para medir el error de clasificación de un árbol de clasificación. Existen otras mediciones que se destacan por ser derivables y por ende, dóciles al momento de optimizar como

- Índice de Gini

  Se define al Índice de Gini para una variable respuesta con K clases de la siguiente manera (James et al., 2013, p.336)

  $$G = \sum_{j=1}^{K} \widehat{p}_{mj}(1 - \widehat{p}_{mj}) \tag{2.58}$$

  Denominado como índice de impureza de nodos, puesto que, si el índice es pequeño significa que $\widehat{p}_{mj}$ es cercano a los extremos: 0 y 1, y por ende, el nodo se inclina considerablemente por una clase haciéndolo un nodo puro.

- Entropía

  Se define a la Entropía para una variable respuesta con K clases de la siguiente manera (James et al., 2013, p.336)

  $$E = -\sum_{j=1}^{K} \widehat{p}_{mj} \log(\widehat{p}_{mj}) \tag{2.59}$$

En la Figura 2.10, lado izquierdo, se observa un árbol de clasificación aplicado al conjunto de datos `DatosCancer` utilizando a $\mathtt{x}_4$ como única variable explicativa. Se parte de un nodo inicial, donde se encuentran todas las observaciones, y se van creando biparticiones hasta llegar a los nodos finales.
Las biparticiones se definen al escoger un punto de corte creando subregiones. Este proceso se puede aplicar iterativamente a cada subregión hasta definir un punto de cese y se le denomina como *partición binaria iterativa*. El punto de corte puede ser tan específico que cada observación tenga su propia región, esto hará que el árbol sea muy profundo. En general, se puede determinar la cantidad de observaciones presentes en cada nodo final, entre más observaciones menos profundidad. En la Figura 2.10 se puede observar este fenómeno.

Ya que el Índice de Gini es derivable, se puede utilizar para determinar los puntos de corte. Se define al Índice de Gini para el nodo m de la siguiente manera (Cortés, 2022, p.37)

$$G_m = N_{izq}\left[\sum_{j=1}^{k} \widehat{p}_{m_{izqj}}(1-\widehat{p}_{m_{izqj}})\right] + N_{der}\left[\sum_{j=1}^{k} \widehat{p}_{m_{derj}}(1-\widehat{p}_{m_{derj}})\right] \tag{2.60}$$

Con m_{der} y m_{izq} los nodos derecho e izquierdo resultantes de la partición, respectivamente, y N_{der}, N_{izq} como su cardinalidad. Se calcula G_m para cada partición posible y se elige un punto de corte que minimice al Índice de Gini, obteniendo nodos lo más puros posibles.

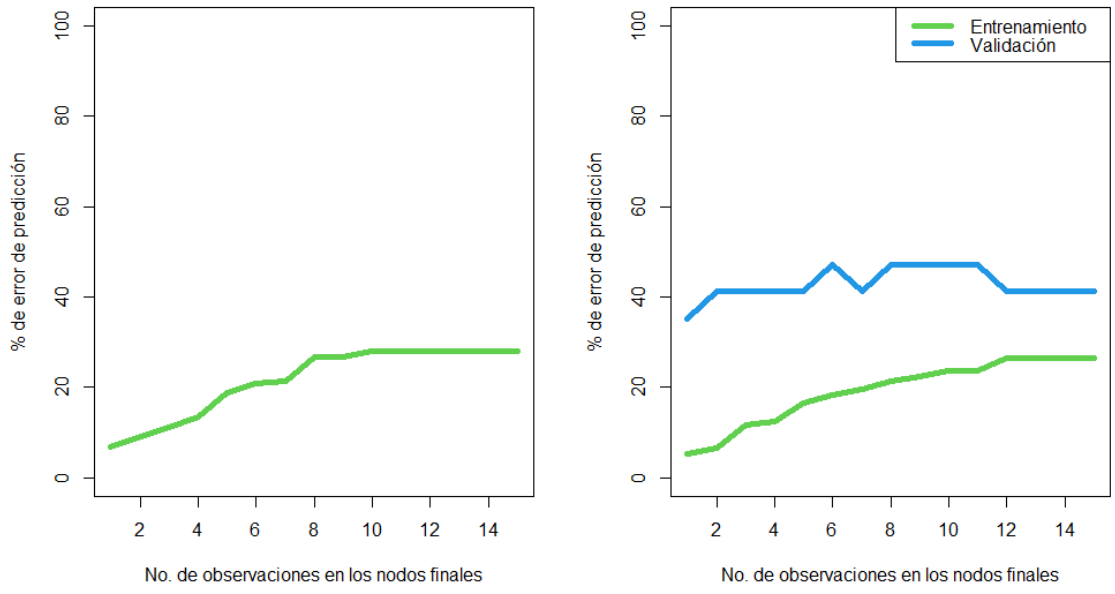

Figura 2.9: Errores de predicción con respecto al número de observaciones en el nodo final del conjunto de datos `DatosCancer`. En este modelo se utilizaron a todas las variables explicativas, en la gráfica del lado derecho se particionó al conjunto de datos en 90 % entrenamiento y 10 % validación.

El árbol de clasificación de la Figura 2.10, superior, fue optimizado utilizando el Índice de Gini. Se observa un primer punto de corte $\mathtt{x}_4 \geq 90$ y $\mathtt{x}_4 < 90$.
En el árbol se pueden observar los 10 nodos finales. Cada nodo expresa la proporción de observaciones de clase 1 dentro de la región, por ejemplo, las observaciones en el primer nodo, de izquierda a derecha, son 29 % de clase 1 y 71 % de clase 0, clasificándolo como `Y`=0, el valor porcentual presente en cada nodo indica el porcentaje del total de observaciones presentes. El árbol superior de la Figura 2.10 obtuvo 32.25 % de error aparente y el árbol inferior un 27.95 %.
En la Figura 2.9 se particionó al conjunto de datos en 90 % entrenamiento y 10 % validación con el objetivo de ser consistentes en las particiones utilizadas en cada método de clasificación, puesto que la función `glm()` no lograba ajustar los parámetros de la regresión logística si se utilizaba un porcentaje menor para el conjunto de entrenamiento. La Figura 2.9 muestra que, cuando se utilizan todos los datos disponibles como entrenamiento, la tasa de error de predicción disminuye entre menos observaciones se tengan en el nodo final, es decir, entre más complejo sea el árbol de clasificación. Cuando se hace una partición al conjunto de datos, en entrenamiento y validación, no se obtiene el mismo resultado. Si se obtiene una mejor predicción para las observaciones de entrenamiento pero una peor para las observaciones de validación ocurre *overfitting* en el modelo.

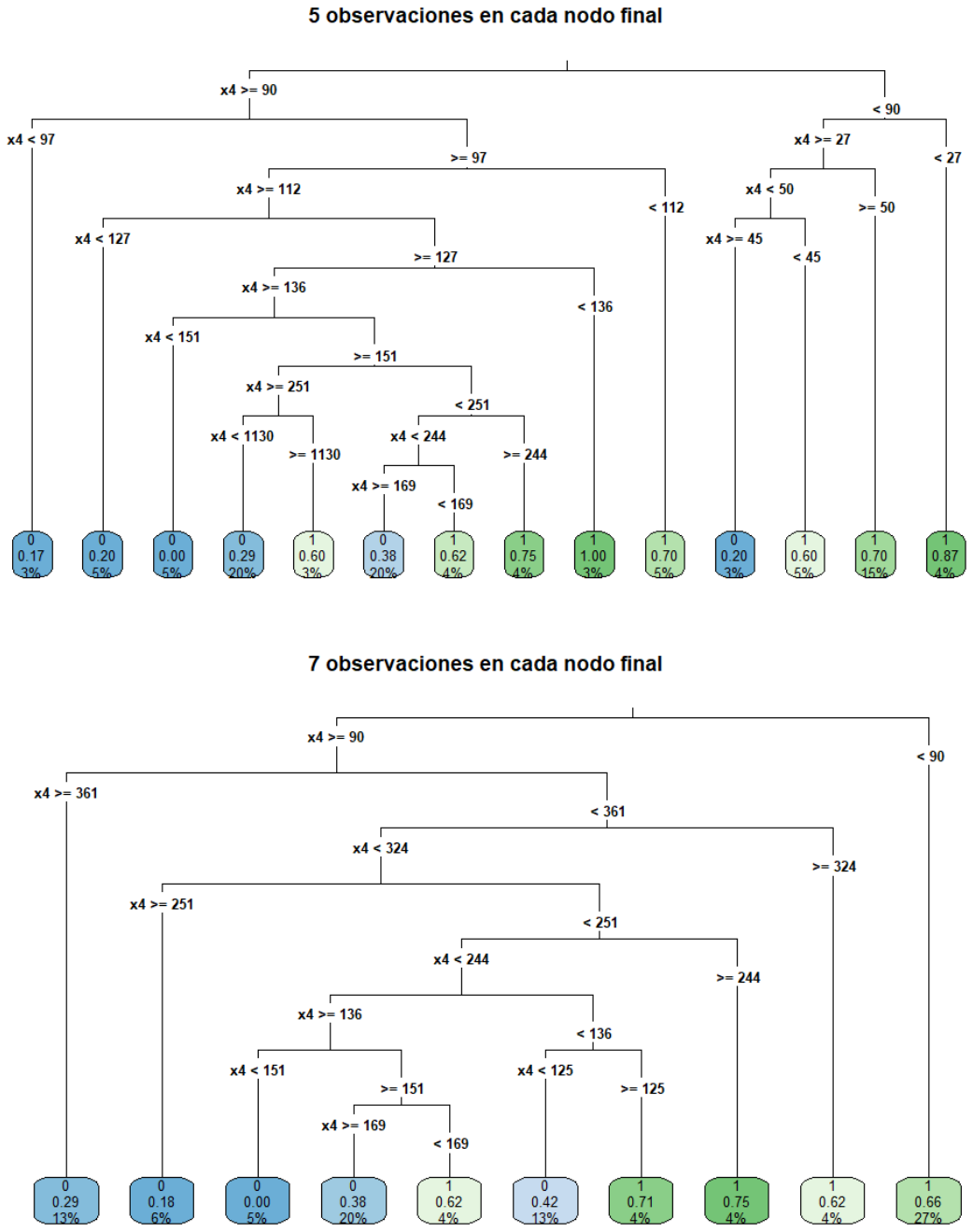

Figura 2.10: Árboles de Clasificación aplicados al conjunto de datos `DatosCancer` utilizando a $\mathtt{x}_4$ como única variable explicativa. Para el árbol inferior se pidió que cada nodo final tuviera un mínimo de 7 observaciones y para el superior, un mínimo de 5 observaciones.

2.7.2. *Bootstrap*

Suponga que se tiene un conjunto de n observaciones $D = \{(x_i, y_i) : i \in 1, \ldots, n\}$ y un parámetro de interés θ. La idea de *Bootstrap* es generar B muestras aleatorias con reemplazo, de tamaño n, del conjunto de datos D, evaluar el estimador de interés $\widehat{\theta}$ con cada muestra obteniendo $\{\widehat{\theta}_1^*, \ldots, \widehat{\theta}_B^*\}$ y estimar algún parámetro de la distribución, por ejemplo, su varianza (Hastie et al., 2009, p.249)

$$\widehat{\mathrm{Var}}(\widehat{\theta}^*) = \frac{1}{B-1}\sum_{i=1}^{B}\left(\widehat{\theta}_i^* - \frac{1}{B}\sum_{i=1}^{B}\widehat{\theta}_i^*\right)^2 \tag{2.61}$$

Al tener $\widehat{\mathrm{Var}}(\widehat{\theta}^*)$ es posible construir un intervalo de $100(1-\alpha)\,\%$ de confianza para el parámetro $\widehat{\theta}$. Los puntos extremos son los siguientes (Efron & Hastie, 2021, p.181)

$$\widehat{\theta} \pm z_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\widehat{\theta}^*)}$$

Cada observación $(x_i, y_i) \in D$ tiene una probabilidad de $1/n$ de ser escogida en cada extracción. Una muestra por *bootstrap* contiene, en promedio, al $63.2\,\%$ de las observaciones originales, esto por lo siguiente

$$P(\text{observación i} \in \text{muestra } \textit{bootstrap} \text{ b}) = 1 - \left(1 - \frac{1}{n}\right)^n, \tag{2.62}$$

y considerando el límite cuando $n \to \infty$

$$\lim_{n\to\infty} 1 - \left(1 - \frac{1}{n}\right)^n = 1 - e^{-1} \approx 0.632.$$

El sesgo estimado usando *bootstrap* es

$$\frac{1}{B}\sum_{i=1}^{B}\widehat{\theta}_i^* - \widehat{\theta}$$

Considere al conjunto de datos `DatosCancer`. La mediana para la variable $\mathtt{x}_1$ es 47. Haciendo un muestreo por *bootstrap* con $B = 1,000$ se obtiene una mediana de 46.57 y una desviación estándar de 0.99. El intervalo del $95\,\%$ de confianza para la mediana es (45.04,48.95). La Figura 2.11, lado izquierdo, expone el histograma de la variable $\mathtt{x}_1$ con su mediana en color rojo y el lado derecho, el histograma de las medianas de las mil muestras con su media en color rojo.

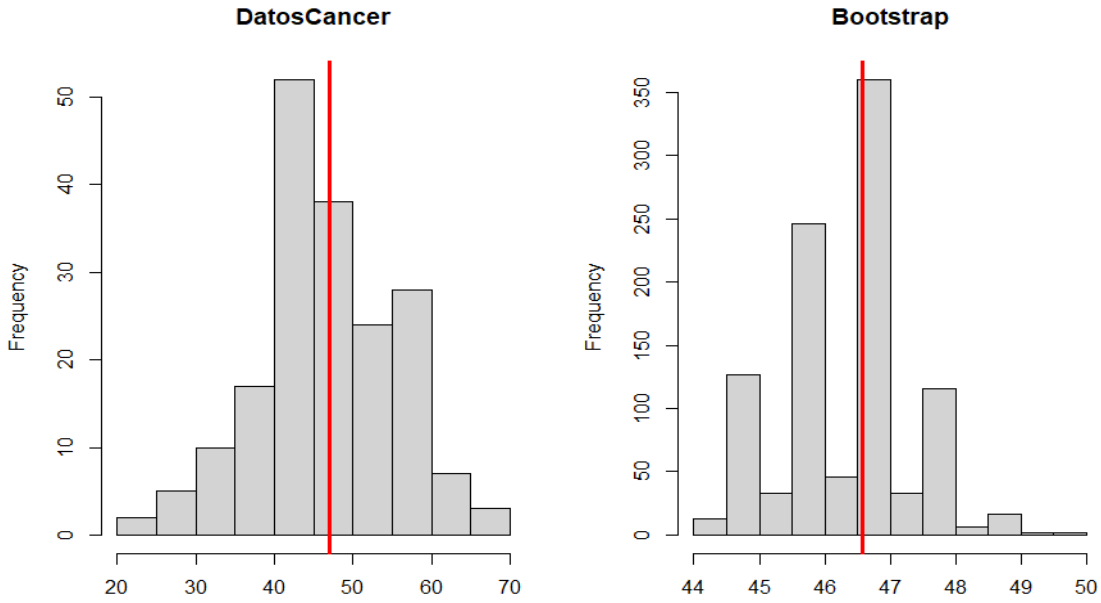

Figura 2.11: Lado izquierdo: histograma del conjunto de datos `DatosCancer` variable $\mathtt{x}_1$ con su mediana en color rojo.
Lado derecho: resultado del muestreo por *bootstrap* de la mediana con B=1,000 con su media en color rojo.

2.7.3. *Bootstrap Aggregation*

El método de *Bootstrap Aggregation* o *Bagging* disminuye la varianza que existe en los árboles de decisión, puesto que, si se parte el conjunto de datos de forma aleatoria en dos subconjuntos y se le aplica un árbol de decisión a cada subconjunto, se obtendrán resultados bastante diferentes (James et al., 2013, p.340).

Si se tienen n observaciones independientes con varianza σ^2, la varianza de la media de las observaciones será σ^2/n. *Bagging* promedia un conjunto de observaciones para reducir la varianza.

Al no contar con una cantidad significativa de datos, se generan B conjuntos de datos utilizando el remuestreo *bootstrap*. Una vez teniendo las B muestras, se aplica un árbol de decisión a cada una y finalmente se promedian la predicciones obtenidas. Se define al método de *bagging*, cuando se tiene una variable respuesta continua, de la siguiente manera (James et al., 2013, p.341)

$$\widehat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B}\widehat{f}^b(x) \tag{2.63}$$

Con $\widehat{f}^b(x)$ el valor estimado para la observación x en la b-ésima muestra.

Para el caso de una variable respuesta categórica se procede de manera similar: para una observación x se toma en cuenta la clase predicha por cada uno de los B

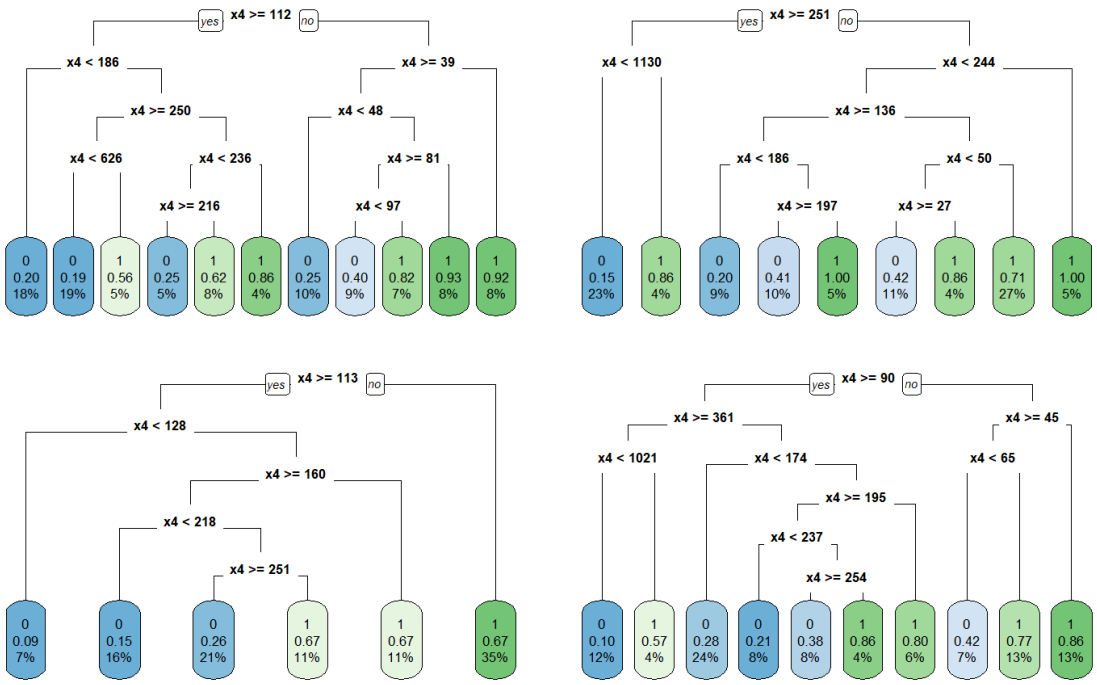

Figura 2.12: Método de *bagging* con $B = 4$ aplicado al conjunto de datos `DatosCancer` con $\mathtt{x}_4$ como única variable explicativa. Se observa que cada árbol de decisión es diferente.

árboles y finalmente se le clasifica a la clase que más se le asignó dentro de las B predicciones.

El método de *bagging* mejora la precisión puesto que se toman en cuenta varios árboles en un mismo proceso. Para seleccionar el valor de B se puede crear una malla con diferentes valores, evaluar el error de predicción para cada uno y seleccionar el que arroje el menor error.

En la Figura 2.12 se pueden observar 4 iteraciones de *bagging* notando que, aunque se parta del mismo conjunto de datos, el resultado es muy diferente y comprueba la alta varianza en los árboles de decisión.

2.7.4. *Random Forest*

El método de *Random Forest* mejora el poder predictivo, con respecto a los árboles de decisión, a cambio de la visualización e interpretación. En *random forest* se limita a que cada nodo utilice un subconjunto de variables explicativas, es decir, del conjunto de p variables explicativas $\mathbf{x}$ se selecciona una muestra aleatoria de m variables. El valor de m sugerido es (Efron & Hastie, 2021, p.327)

$$m = \begin{cases} \sqrt{p} & \text{clasificación } (\mathtt{Y} \text{ discreta}) \\ \frac{p}{3} & \text{regresión } (\mathtt{Y} \text{ continua}) \end{cases}$$

Random Forest utiliza a los siguiente parámetros:

- *ntree*: número de árboles.

  Este número depende del conjunto de datos D_b y se busca una convergencia en el error de predicción. Usualmente se necesita un valor de 100 para asegurar la convergencia (James et al., 2013, p.341).

- *mtry*: número de variables a utilizar en los nodos.

El método de *random forest* se compone de

1. Se define el valor de *ntree*.
2. Para $j = 1$ hasta $j =$*ntree*:
   - Se genera una muestra por *bootstrap* D_b de tamaño n del conjunto de datos D.
   - Se ajusta un árbol de decisión para cada muestra generada D_b y para cada nodo:
     - Se define el valor de *mtry*, es decir, las variables a utilizar en el nodo.
     - Se selecciona a la mejor variable entre las seleccionadas y se define su punto de corte.
     - Se definen a los nuevos nodos, derecho e izquierdo, hasta alcanzar la profundidad deseada del árbol.
3. Guardar la colección de los *ntree* árboles de decisión.
4. Finalmente, como fue expuesto en *bagging*, para el caso donde se busca hacer una clasificación de la observación x, se considera el promedio de la clase más predicha para la observación en los *ntree* árboles de decisión.

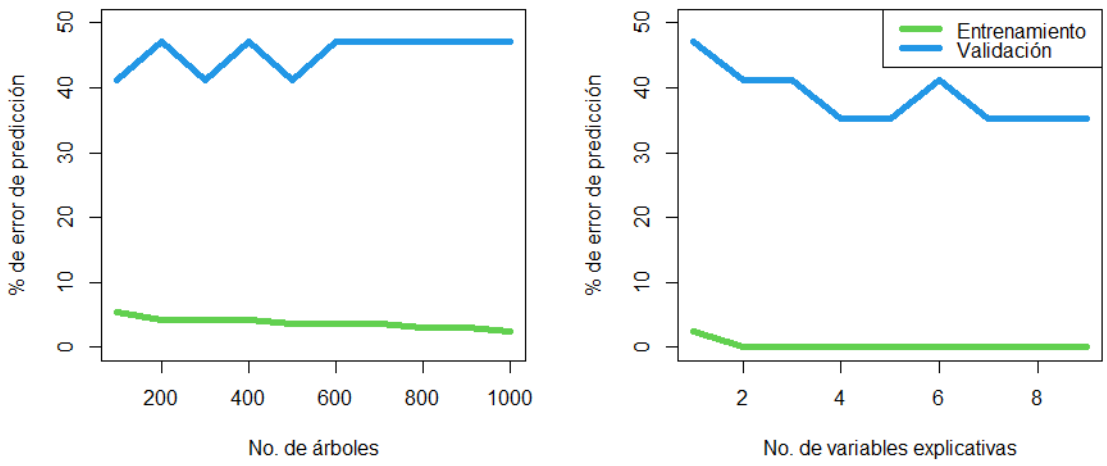

Figura 2.13: Comparación entre la tasa de error y el número de variables explicativas y el número de árboles. Se utilizó el conjunto de datos `DatosCancer` y se particionó una vez al conjunto en 90 % entrenamiento y 10 % validación.

Random Forest se compone de un número *ntree* de árboles distintos. En la Figura 2.13, lado izquierdo, se puede apreciar la diferencia en las tasas de error conforme el número de árboles aumenta utilizando una sola variable explicativa. La tasa de error de validación se mantiene entre 40-50 % y la de entrenamiento es menor al 10 %. La gráfica del lado derecho describe el comportamiento de la tasa de error conforme el número de variables aumenta considerando únicamente 1,000 árboles. De la Figura 2.13 se puede concluir que, para el conjunto de datos `DatosCancer`, el número de árboles no influye en la tasa de error de validación de la misma forma que lo hace el número de variables explicativas utilizadas.

2.8. Distribución Gaussiana

2.8.1. Distribución Gaussiana Multivariada

La distribución Gaussiana o distribución Normal, es un distribución utilizada en las variables continuas. Para el caso de una variable $\mathtt{x}$ la función de densidad es

$$f(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right] \tag{2.64}$$

Donde $x \in \mathbb{R}$, μ es la media y σ^2 la varianza.
Para un vector de p variables $\mathbf{x}^\top = (\mathtt{x}_1, \ldots, \mathtt{x}_p)$, la función de densidad es

$$f(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right] \tag{2.65}$$

Con $x \in \mathbb{R}^p$, μ el vector de medias de p dimensiones y Σ la matriz de covarianzas.
La matriz inversa de Σ es conocida como *matriz de concentración* y su expresión es

$$\Lambda \equiv \Sigma^{-1} \tag{2.66}$$

Una propiedad de la distribución Gaussiana es que, si la unión de dos conjuntos de variables tiene distribución Gaussiana, entonces las distribuciones de un conjunto condicionado al otro y las marginales de cada conjunto son Gaussianas (Bishop, 2006, p.85).

Suponga un vector $\mathbf{x}$ de p variables continuas con distribución Gaussiana $N(\mu, \Sigma)$, se divide a $\mathbf{x}$ en dos subconjuntos $\mathbf{x}_a$ y $\mathbf{x}_b$ tal que las primeras N variables de $\mathbf{x}$ están contenidas en $\mathbf{x}_a$ y en $\mathbf{x}_b$, las $p-N$ restantes. El vector $\mathbf{x}$ está expresado de la siguiente manera

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix}$$

Dada la partición de $\mathbf{x}$, el vector de medias μ se expresa como

$$\mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}$$

Donde μ_a es el vector de medias de $\mathbf{x}_a$ y μ_b el de $\mathbf{x}_b$.
La matriz de covarianzas Σ queda expresada como

$$\Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}$$

Donde Σ_{aa} y Σ_{bb} corresponden a las matrices de covarianzas dentro de las variables de $\mathbf{x}_a$ y $\mathbf{x}_b$, respectivamente y Σ_{ba}, Σ_{ab} a las matrices de covarianzas entre las variables de $\mathbf{x}_a$ y $\mathbf{x}_b$.
Dada la propiedad de simetría presente en Σ, se concluye que Σ_{aa} y Σ_{bb} son simétricas y $\Sigma_{ba} = \Sigma_{ab}^\top$.
La expresión de la matriz de concentración, dada la partición de $\mathbf{x}$, es

$$\Lambda = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix}$$

Dado que $\Lambda \equiv \Sigma^{-1}$ y la matriz inversa de una matriz simétrica es simétrica, se concluye que Λ_{aa} y Λ_{bb} son simétricas y $\Lambda_{ab}^\top = \Lambda_{ba}$, inclusive (Bishop, 2006, p.87)

$$\begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}^{-1} = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix}$$

La distribución condicional de $\mathbf{x}_a|\mathbf{x}_b$ es de la siguiente manera (Bishop, 2006, p.90)

$$\mathbf{x}_a|\mathbf{x}_b \sim N(\mu_{a|b}, \Lambda_{aa}^{-1}) \tag{2.67}$$

Donde

$$\mu_{a|b} = \mu_a - \Lambda_{aa}^{-1}\Lambda_{ab}(\mathbf{x}_b - \mu_b) \tag{2.68}$$

La distribución marginal de $\mathbf{x}_a$ es (Anderson, 2009, p.26)

$$\mathbf{x}_a \sim N(\mu_a, \Sigma_{aa}) \tag{2.69}$$

Donde μ_a es el vector de medias de las variables en $\mathbf{x}_a$ y Σ_{aa} la matriz de covarianza entre ellas.

Considere a las variables continuas de la Figura 1.3 con la que se simularon los datos de la Sección 1.2. Para la clase 0, las variables $(\mathbf{x}_1, \ldots, \mathbf{x}_6) \sim N(\mu, \Sigma)$ con $\mu = \mathbf{0}$ y Σ expresada de la siguiente manera

$$\Sigma = \begin{pmatrix} 1.000 & 0.300 & 0.090 & 0.027 & 0.008 & 0.002 \\ 0.300 & 1.000 & 0.300 & 0.090 & 0.027 & 0.008 \\ 0.090 & 0.300 & 1.000 & 0.300 & 0.090 & 0.027 \\ 0.027 & 0.090 & 0.300 & 1.000 & 0.300 & 0.090 \\ 0.008 & 0.027 & 0.090 & 0.300 & 1.000 & 0.300 \\ 0.002 & 0.008 & 0.027 & 0.090 & 0.300 & 1.000 \end{pmatrix} \tag{2.70}$$

Con matriz de concentración

$$\Lambda = \begin{pmatrix} 1.099 & -0.330 & 0 & 0 & 0 & 0 \\ -0.330 & 1.198 & -0.330 & 0 & 0 & 0 \\ 0 & -0.330 & 1.198 & -0.330 & 0 & 0 \\ 0 & 0 & -0.330 & 1.198 & -0.330 & 0 \\ 0 & 0 & 0 & -0.330 & 1.198 & -0.330 \\ 0 & 0 & 0 & 0 & -0.330 & 1.099 \end{pmatrix} \tag{2.71}$$

Suponga la partición del vector de variables $\mathbf{x}^\top = (\mathtt{x}_1, \mathtt{x}_2, \ldots, \mathtt{x}_6)$ en $\mathbf{x}_\alpha^\top = (\mathtt{x}_1, \mathtt{x}_2, \mathtt{x}_3)$ y $\mathbf{x}_\beta^\top = (\mathtt{x}_4, \mathtt{x}_5, \mathtt{x}_6)$, entonces $\mu_\alpha = \mu_\beta = (0, 0, 0)^\top$ y el vector de medias μ es el siguiente

$$\mu = \begin{pmatrix} \mu_\alpha \\ \mu_\beta \end{pmatrix}$$

La matriz de covarianzas queda expresada de la siguiente manera

$$\Sigma = \begin{pmatrix} \Sigma_{\alpha\alpha} & \Sigma_{\alpha\beta} \\ \Sigma_{\beta\alpha} & \Sigma_{\beta\beta} \end{pmatrix}$$

Con

$$\Sigma_{\alpha\alpha} = \Sigma_{\beta\beta} = \begin{pmatrix} 1.000 & 0.300 & 0.090 \\ 0.300 & 1.000 & 0.300 \\ 0.090 & 0.300 & 1.000 \end{pmatrix}$$

Y

$$\Sigma_{\alpha\beta}^\top = \Sigma_{\beta\alpha} = \begin{pmatrix} 0.027 & 0.090 & 0.300 \\ 0.008 & 0.027 & 0.090 \\ 0.002 & 0.008 & 0.027 \end{pmatrix}$$

Suponga que $\mathbf{x}_\beta^\top = (x_4, x_5, x_6)$, el vector de media $\mu_{\alpha|\beta}$ es

$$\begin{aligned} \mu_{\alpha|\beta} &= \mu_\alpha - \Lambda_{\alpha\alpha}^{-1}\Lambda_{\alpha\beta}(\mathbf{x}_\beta - \mu_\beta) \\ &= \mathbf{0} - \begin{pmatrix} 1.099 & -0.330 & 0 \\ -0.330 & 1.198 & -0.330 \\ 0 & -0.330 & 1.198 \end{pmatrix}^{-1} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ -0.330 & 0 & 0 \end{pmatrix} (x_4, x_5, x_6)^\top \\ &= \mathbf{0} - \begin{pmatrix} 1.000 & 0.300 & 0.090 \\ 0.300 & 1.000 & 0.300 \\ 0.090 & 0.300 & 1.000 \end{pmatrix} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ -0.330 & 0 & 0 \end{pmatrix} (x_4, x_5, x_6)^\top \\ &= \mathbf{0} - \begin{pmatrix} -0.029 & 0 & 0 \\ -0.099 & 0 & 0 \\ -0.330 & 0 & 0 \end{pmatrix} (x_4, x_5, x_6)^\top = (0.029x_4, 0.099x_4, 0.330x_4)^\top \end{aligned} \tag{2.72}$$

La función de densidad de $\mathbf{x}_\alpha|\mathbf{x}_\beta$ es

$$f_{\mathbf{x}_\alpha|\mathbf{x}_\beta}(x_\alpha) = \frac{1}{(2\pi)^3|\Lambda_{\alpha\alpha}^{-1}|^{1/2}} \exp\left\{ -\frac{1}{2}[x_\alpha^\top - (0.029x_4, 0.099x_4, 0.330x_4)]\Lambda_{\alpha\alpha} \right. \\ \left. [x_\alpha - (0.029x_4, 0.099x_4, 0.330x_4)^\top] \right\}$$

Con $x_\alpha^\top = (x_1, x_2, x_3)$ y

$$\Lambda_{\alpha\alpha}^{-1} = \begin{pmatrix} 1.000 & 0.300 & 0.090 \\ 0.300 & 1.000 & 0.300 \\ 0.090 & 0.300 & 1.000 \end{pmatrix}$$

Si $\mathbf{x}_\beta^\top = (0.6, 0.7, -2.4)$, entonces $\mathbf{x}_\alpha|\mathbf{x}_\beta \sim N[(0.017, 0.059, 0.198)^\top, \Lambda_{\alpha\alpha}^{-1}]$ con función de densidad

$f_{\mathbf{x}_\alpha|\mathbf{x}_\beta}(x_\alpha) = \frac{1}{(2\pi)^3|\Lambda_{\alpha\alpha}^{-1}|^{1/2}} \exp\left\{-\frac{1}{2}[x_\alpha^\top - (0.017, 0.059, 0.198)]\Lambda_{\alpha\alpha}[x_\alpha - (0.017, 0.059, 0.198)^\top]\right\}$

2.8.2. Distribución Gaussiana Condicional

La distribución Gaussiana Condicional está definida para una mezcla entre variables discretas y continuas. El conjunto de variables $V = \triangle \cup \Gamma$ es dividido en el conjunto de d variables discretas $\triangle$ y en el de r variables continuas Γ, entonces el vector de variables ($\mathbf{i}$,$\mathbf{x}$) con $\mathbf{i}$ el vector de variables binarias y $\mathbf{x}$ el de variables continuas es visualizado de la siguiente manera (Lauritzen, 1996, p.158)

$$(\mathbf{i}, \mathbf{x}) = (\mathtt{i}_1, \ldots, \mathtt{i}_d, \mathtt{x}_1, \ldots, \mathtt{x}_r)$$

En este trabajo se abordará el caso donde $\mathbf{i}$ es un vector binario. Como $\mathbf{i}$ es binario con d entradas, tiene 2^d posibles estados o celdas i_s. La distribución marginal de $\mathbf{i}$ está expresada en términos del coeficiente γ_s tal que

$$p(\mathbf{i} = i_s) = \gamma_s > 0$$

Con $s \in \{1, \ldots, 2^d\}$ y

$$\sum_{s=1}^{2^d} \gamma_s = 1$$

La distribución condicional de $\mathbf{x}$ dado un valor particular de $\mathbf{i}$ es una Gaussiana (Lauritzen, 1996, p.159)

$$f(\mathbf{x}|\mathbf{i} = i_s) = N(\mathbf{x}|\mu(i_s), \Sigma(i_s))$$

Donde $\mu(i_s)$ es el vector de medias y $\Sigma(i_s)$ la matriz de covarianzas que dependen del valor i_s.
La distribución marginal de $\mathbf{x}$ se obtiene al sumar la función de densidad conjunta sobre todos los 2^d posibles valores de $\mathbf{i}$ obteniendo (Bishop, 2006, p.431)

$$f(\mathbf{x}) = \sum_{s=1}^{2^d} p(\mathbf{i}) f(\mathbf{x}|\mathbf{i}) = \sum_{s=1}^{2^d} \gamma_s N(\mathbf{x}|\mu(i_s), \Sigma(i_s))$$

Finalmente, la función de densidad conjunta $f(\mathbf{i}, \mathbf{x})$ es la siguiente (Bishop, 2006, p.431)

$$f(\mathbf{i}, \mathbf{x}) = p(\mathbf{i}) f(\mathbf{x}|\mathbf{i}) = \gamma_s N(\mathbf{x}|\mu(i_s), \Sigma(i_s))$$

La gráfica de la Figura 1.3 corresponde a una distribución Gaussiana Condicional con **i** el vector de variables binarias y **x** el vector de variables continuas. La gráfica muestra que las variables continuas son independientes de las binarias, por lo que $\mu(i_s) = \mathbf{0}$ y $\Sigma(i_s) = \Sigma_k$ para todo i_s, es decir

$$f(\mathbf{x}) = \sum_{s=1}^{16} \gamma_s N(\mathbf{x}|\mathbf{0}, \Sigma_k) = N(\mathbf{x}|\mathbf{0}, \Sigma_k) \sum_{s=1}^{16} \gamma_s = N(\mathbf{x}|\mathbf{0}, \Sigma_k)$$

En la Tabla 1.5 se pueden observar a las diferentes distribuciones Gaussianas de **x** dado el valor de **i**. Ya que **x** es independiente de **i**, mantiene la misma distribución Gaussiana sin importar el valor de **i** y únicamente cambia con respecto a la clase k. Para la clase 0, $\mathbf{x} \sim N(\mathbf{0}, \Sigma_0)$ y para la clase 1, $\mathbf{x} \sim N(\mathbf{0}, \Sigma_1)$.

Suponga a la clase 0 y una observación $(i, x) = (1, 0, 0, 0, 0, 0.9, 2, 1.5, -0.3, -1, 0.2)$. En este caso, al suponer una clase k, la distribución marginal de **i** está expresada de la siguiente manera

$$p_k(\mathbf{i} = i_s) = p(\mathbf{i} = i_s | K = k) = \gamma_s^k$$

La función de densidad conjunta es

$$\begin{aligned} f(i, x) &= p_0(i) f(x|i) \\ &= \frac{p_0(1,0)p_0(0,0)p_0(0,0)}{p_0(0)p_0(0)} f(0.9) \prod_{i=2}^{6} f(\mathtt{x}_i|\mathtt{x}_{i-1}) \\ &= \frac{0.175 \times 0.325^2}{0.5^2} \times 0.81 \times \prod_{i=2}^{6} f(\mathtt{x}_i|\mathtt{x}_{i-1}) \end{aligned}$$

Para obtener $f(\mathtt{x}_i|\mathtt{x}_{i-1})$ se aplica lo visto en (2.67) y (2.68). Como las variables del vector **x** están correlacionadas en forma de cadena, únicamente se utiliza de (2.70) a la submatriz correspondiente a $\mathtt{x}_i$ y $\mathtt{x}_{i-1}$, con $i \in \{2, \ldots, 6\}$, la cual es

$$\begin{array}{c} \\ \mathtt{x}_{i-1} \\ \mathtt{x}_i \end{array} \begin{array}{c} \begin{array}{cc} \mathtt{x}_{i-1} & \mathtt{x}_i \end{array} \\ \begin{pmatrix} 1.0 & 0.3 \\ 0.3 & 1.0 \end{pmatrix} \end{array}$$

Con matriz de concentración

$$\begin{array}{c} \\ \mathtt{x}_{i-1} \\ \mathtt{x}_i \end{array} \begin{array}{c} \begin{array}{cc} \mathtt{x}_{i-1} & \mathtt{x}_i \end{array} \\ \begin{pmatrix} 1.09 & -0.32 \\ -0.32 & 1.09 \end{pmatrix} \end{array}$$

El vector de media $\mu_{\mathtt{x}_i|\mathtt{x}_{i-1}}$ es

$$\begin{aligned} \mu_{\mathtt{x}_i|\mathtt{x}_{i-1}} &= \mu_{\mathtt{x}_i} - \Lambda^{-1}_{\mathtt{x}_i\mathtt{x}_i} \Lambda_{\mathtt{x}_i\mathtt{x}_{i-1}} (x_{i-1} - \mu_{\mathtt{x}_{i-1}}) \\ &= 0 - 1.09^{-1} \times (-0.32) \times (x_{i-1}) \\ &= 0.3 x_{i-1} \end{aligned} \tag{2.73}$$

Concluyendo que $\mathtt{x}_i|\mathtt{x}_{i-1} \sim N(0.3x_{i-1}, 0.91)$ con función de densidad

$$f(x_i|x_{i-1}) = \frac{1}{(2\pi \times 0.91)^{1/2}} \exp\left[-\frac{1}{2 \times 0.91}(x_i - 0.3x_{i-1})^2\right] \tag{2.74}$$

Finalmente

$$\begin{aligned} f(i,x) &= \frac{0.175 \times 0.325^2}{0.5^2} \times 0.81 \times \prod_{i=2}^{6} f(\mathtt{x}_i|\mathtt{x}_{i-1}) \\ &= 0.07 \times 0.81 \times 0.97 \times 0.83 \times 0.20 \times 0.15 \times 0.70 \\ &= 0.001 \end{aligned}$$

En el siguiente capítulo se medirá el poder predictivo, de cada uno de los métodos mencionados en este capítulo, para los dos conjuntos de datos mencionados: `DatosCancer` y `DatosSim`.

Capítulo 3

Aplicación de los Métodos de Clasificación

Se reportó la media de los errores obtenidos en cada método con mil repeticiones utilizando el método de *validation set approach*. Ya que se considera la media de un conjunto de mil elementos, la desviación estándar de la media está descrita de la siguiente manera:

$$\text{SE} = \sigma_{\bar{\mathtt{x}}} = \frac{\sigma}{\sqrt{1000}}$$

Se expone la paquetería utilizada y el código empleado en **R** en los Anexos A y B, respectivamente. Se utilizó la notación de **R**, que está basada en la notación de Wilkinson, para expresar al modelo aditivo y al de interacciones dos a dos.

3.1. Estudio de Cáncer de Mama

Se reportaron los errores obtenidos partiendo el conjunto de datos en 90 % entrenamiento y 10 % validación, esta configuración fue seleccionada puesto que la función `glm()` no lograba ajustar los parámetros si se utilizaba un porcentaje menor para el conjunto de entrenamiento. De igual forma, se utilizó la opción `singular.ok=TRUE` en la función `glm()` puesto que en algunas particiones del conjunto de datos el algoritmo no convergía.

Se aplicaron los métodos a dos modelos: al modelo aditivo (3.1)

$$\mathtt{Y} \sim \mathtt{x}_1 + \mathtt{x}_2 + \mathtt{x}_3 + \mathtt{x}_4 + \mathtt{x}_5 + \mathtt{x}_6 + \mathtt{x}_7 + \mathtt{x}_8 + \mathtt{x}_9 \tag{3.1}$$

Y al modelo con interacciones dos a dos (3.2)

$$\mathtt{Y} \sim \mathtt{x}_1 + \mathtt{x}_2 + \mathtt{x}_3 + \mathtt{x}_4 + \mathtt{x}_5 + \mathtt{x}_6 + \mathtt{x}_7 + \mathtt{x}_8 + \mathtt{x}_9 + \sum_{i=1}^{8} \sum_{j=i+1}^{9} \mathtt{x}_i \mathtt{x}_j \tag{3.2}$$

3.1.1. Regresión Logística

Se ajustó un modelo de regresión logística al modelo aditivo (3.1) y al modelo con interacciones dos a dos (3.2), clasificando a una observación x a la clase 1 si $P(\mathtt{Y} = 1|\mathbf{x} = x) \geq 0.5$ y clase 0 en otro caso. Se obtuvo una tasa de error de validación global de 38.92 % para el modelo aditivo y 40.35 % para el modelo con interacciones dos a dos. En la Tabla 3.1 se exponen los errores por clase y la desviación estándar.

Tabla 3.1: Errores de validación (%) y desviación estándar de la media del error global al ajustar una regresión logística al modelo aditivo y al modelo con interacciones dos a dos. Se particionó al conjunto de datos en 90 % entrenamiento y 10 % validación y se repitió mil veces.

	Error de Validación			
Modelo	Global	Clase 0	Clase 1	SE
Regresión logística aditiva	38.92	36.82	41.30	0.36
Regresión con interacciones 2 a 2	40.35	37.10	44.01	0.36

Nota: El modelo regresión logística aditiva corresponde al modelo aditivo (3.1).

3.1.2. *Stepwise Selection*

Se aplicó el algoritmo *backward stepwise selection* al modelo de regresión logística con interacciones dos a dos (3.2) de 46 parámetros utilizando a los criterios AIC y BIC, obteniendo un tasa de error de validación global de 37.15 % para el criterio AIC y 33.25 % para el criterio BIC (véase Tabla 3.2).

3.1.3. Métodos de Regularización

Se regularizó al modelo de regresión logística con interacciones dos a dos (3.2). Para las regularizaciones se utilizó la función `cv.glmnet()` que calcula el valor óptimo de λ utilizando *10-fold cross-validation*. Se obtuvo una tasa de error de validación global de 34.02 % para *ridge regression* y 33.29 % para *lasso*.
Para el caso de *elastic net*, como se expuso en la ecuación (2.37), se necesitó calibrar el parámetro α, obteniendo una tasa de error global de 33.96 %.

Tabla 3.2: Tasas de errores de validación (%), desviación estándar de la media del error global y promedio de número coeficientes de los modelos de regresión logística a la que se le aplicó *stepwise selection* y regularización. Se particionó al conjunto de datos en 90 % entrenamiento y 10 % validación y se repitió mil veces.

Modelo de regresión logística	Error de Validación			SE	Promedio No. Coeficientes
	Global	Clase 0	Clase 1		
Stepwise selection AIC	37.15	33.31	41.48	0.37	18.51
Stepwise selection BIC	33.25	26.18	41.20	0.33	8.59
Ridge Regression	34.02	28.95	39.73	0.34	46.00
Lasso	33.29	28.05	39.18	0.33	20.83
Elastic net	33.96	28.68	39.90	0.33	25.67

Nota: Los modelos anteriores fueron aplicados a la regresión logística con interacciones 2 a 2 (3.2).

3.1.4. Análisis Discriminante

Se aplicó LDA y QDA al modelo aditivo (3.1), obteniendo unas tasas de errores de validación globales de 39.40 % para LDA y 40.61 % en QDA. En la Tabla 3.7 se encuentran los errores de validación por grupo para cada método.

Tabla 3.3: Errores de validación (%) para los métodos de LDA y QDA. Se particionó al conjunto de datos en 90 % entrenamiento y 10 % validación y se repitió mil veces.

Modelo	Error de Validación		
	Global	Clase 0	Clase 1
LDA	39.40	36.18	43.01
QDA	40.61	36.78	44.91

Nota: Ambos modelos fueron aplicados al modelo aditivo (3.1).

3.1.5. *Random Forest*

Se aplicó *random forest* al modelo aditivo (3.1) y se afinó el parámetro *mtry*. El parámetro tiene el siguiente rango: $mtry \in \{1, 2, \ldots, 9\}$. En la Figura 2.13 se pudo notar que el número de variables influía en la tasa de error de predicción de mayor manera que el número de árboles, es por eso que se decidió utilizar 1,000 árboles para medir el poder predictivo.

Finalmente, se obtuvo una tasa de error de validación global de 37.88 % y una desviación estándar de la media del error global de 0.24.

3.1.6. Resultados

En la Tabla 3.4 se exponen las tasas de errores de validación para los métodos expuestos en este trabajo. Se observa que el método con mejor resultado fue la regresión logística con interacciones 2 a 2 a la cual se le aplicó *stepwise selection* utilizando el criterio BIC con 33.25 % de tasa de error de validación. Destaca este resultado puesto que es el método clásico de regresión logística a la que se le aplicó un método de selección de variables, de hecho, este método promedió el menor número de coeficientes entre los modelos de la Tabla 3.2 facilitando la interpretación del modelo y al ser un modelo clásico, tiene el beneficio de poder hacer inferencia estadística.

Los métodos de regularización *lasso, ridge regression* y *elastic net* obtuvieron de las mejores tasas de error de validación entre los métodos, por encima de métodos modernos algorítmicos como *random forest*.

Se esperaría que los métodos modernos algorítmicos dominaran sobre los métodos clásicos, en este caso no fue así y una posible razón podría ser el número muy pequeño de observaciones para un espacio de 10 dimensiones, al contar con más observaciones, se pueden obtener mejores modelos y predicciones.

La Figura 3.1 expone de manera gráfica a los errores de validación, por clase y globales, obtenidos de cada modelo aplicado al conjunto de datos.

Tabla 3.4: Comparación, entre los modelos, de los errores de validación (%) y desviación estándar de la media del error global. Se particionó al conjunto de datos en 90 % entrenamiento y 10 % validación y se repitió mil veces.

	Error de Validación			
Modelo	Global	Clase 0	Clase 1	SE
Regresión Logística aditiva	38.92	36.82	41.30	0.36
Regresión con interacciones 2 a 2	40.35	37.10	44.01	0.36
Regresión con *stepwise selection* AIC	37.15	33.31	41.48	0.37
Regresión con *stepwise selection* BIC	**33.25**	26.18	41.20	0.33
Regresión regularizada por Ridge	34.02	28.95	39.73	0.34
Regresión regularizada por Lasso	**33.29**	28.05	39.18	0.33
Regresión regularizada por E. Net	**33.96**	28.68	39.90	0.33
Análisis discriminante lineal LDA	39.40	36.18	43.01	0.35
Análisis discriminante cuadrático QDA	40.61	36.78	44.90	0.36
Random Forest con 1,000 árboles	37.88	29.17	47.26	0.24

Nota: Los modelos de regresión con interacciones 2 a 2, *stepwise selection* y regularizados fueron aplicados al modelo con interacciones 2 a 2 (3.2). LDA, QDA y *random forest* fueron aplicados al modelo aditivo (3.1).

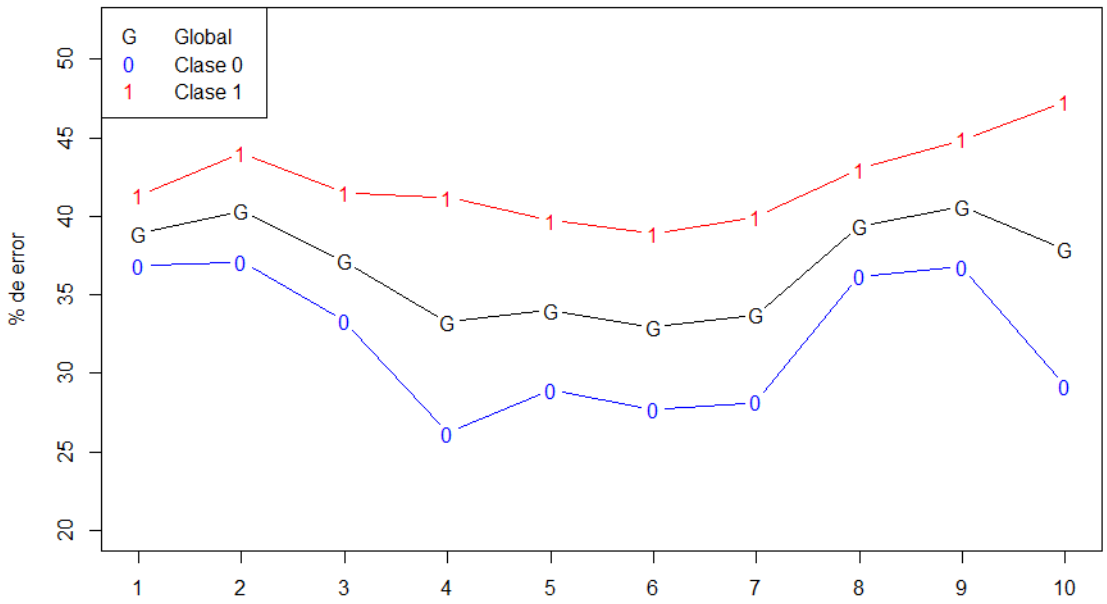

Figura 3.1: Errores de validación (%) por clase y global de cada modelo aplicado al conjunto de datos `DatosCancer`.
Modelos: 1.Regresión logística aditiva 2.Regresión con interacciones 2 a 2 3.*Stepwise selection* AIC 4.*Stepwise selection* BIC 5.*Ridge Regression* 6.*Lasso* 7.*Elastic Net* 8.LDA 9.QDA 10.*Random Forest* con 1,000 árboles.

3.2. Estudio de simulación

Se generaron dos bases de datos a partir de la función de densidad de la distribución Gaussiana Condicional (1.5) con 200 observaciones cada una. Una base de datos se utilizó para entrenamiento y la otra para validación. Cada clase contiene 100 observaciones utilizando $\rho = 0.3$ para la clase 0 y $\rho = -0.3$ para la clase 1.

Obsérvese que en este estudio de simulación las medias, para ambas clases, son cero y la única diferencia entre las clases está en las matrices de covarianzas. Este caso se ilustra en Bartlett & Please (1963).

3.2.1. Regresión Logística

Se ajustó un modelo de regresión logística al modelo aditivo y al modelo con interacciones dos a dos, clasificando a una observación x a la clase 1 si $P(\mathtt{Y} = 1|\mathbf{x} = x) \geq 0.5$ y clase 0 en otro caso. Se obtuvo una tasa de error de validación global de 50.07 % para el modelo aditivo y 29.46 % para el modelo con interacciones dos a dos. En la Tabla 3.5 se exponen los errores de validación por clase y la desviación

estándar.

Tabla 3.5: Errores de validación (%) y desviación estándar de la media del error global al ajustar una regresión logística al modelo aditivo y al modelo con interacciones dos a dos. Se simularon 200 observaciones para entrenamiento y 200 para validación repitiendo el proceso mil veces.

	Error de Validación			
Modelo	Global	Clase 0	Clase1	SE
Regresión logística aditiva	50.07	49.92	50.23	0.11
Regresión con interacciones 2 a 2	29.46	29.30	29.62	0.11

3.2.2. *Stepwise Selection*

Se aplicó el algoritmo *backward stepwise selection* al modelo de regresión logística con interacciones dos a dos de 56 parámetros utilizando a los criterios AIC y BIC, obteniendo un tasa de error de validación global de 28.59 % para el criterio AIC y una media de 32.39 coeficientes. Para el criterio BIC se obtuvo una tasa de error de validación global de 26.63 % y una media de 20.43 coeficientes.

3.2.3. Métodos de Regularización

Se regularizó el modelo de regresión logística con interacciones dos a dos. De igual que en Subsección 3.1.3, se utilizó la función `cv.glmnet()` y se utilizó *10-fold cross-validation* para determinar el valor de λ. Se obtuvo una tasa de error de validación global de 27.61 % para *ridge regression* y 26.74 % para *lasso*.
Para el caso de *elastic net*, se analizó el valor óptimo para el parámetro α. En la Figura 3.2, lado izquierdo, se observa la comparación entre los errores de validación y el valor de α, donde $\alpha = 0.6$ fue el mejor resultado. Definiendo a $\alpha = 0.6$ se obtuvo una tasa de error de validación global de 26.89 %. En la Tabla 3.6 se pueden observar las tasas de errores de validación por clase, desviación estándar de la media del error global y el promedio de coeficientes después de mil repeticiones.

Tabla 3.6: Tasas de errores de validación (%), desviación estándar de la media del error global y promedio del número de coeficientes de los modelos de regularización. Se simularon 200 observaciones para entrenamiento y 200 para validación repitiendo el proceso mil veces.

	Error de Validación				
Regularización	Global	Clase 0	Clase1	SE	Promedio No. Coeficientes
Ridge Regression	27.61	27.31	27.91	0.11	56.00
Lasso	26.74	26.84	26.64	0.11	29.27
Elastic Net	26.89	26.95	26.83	0.11	30.32

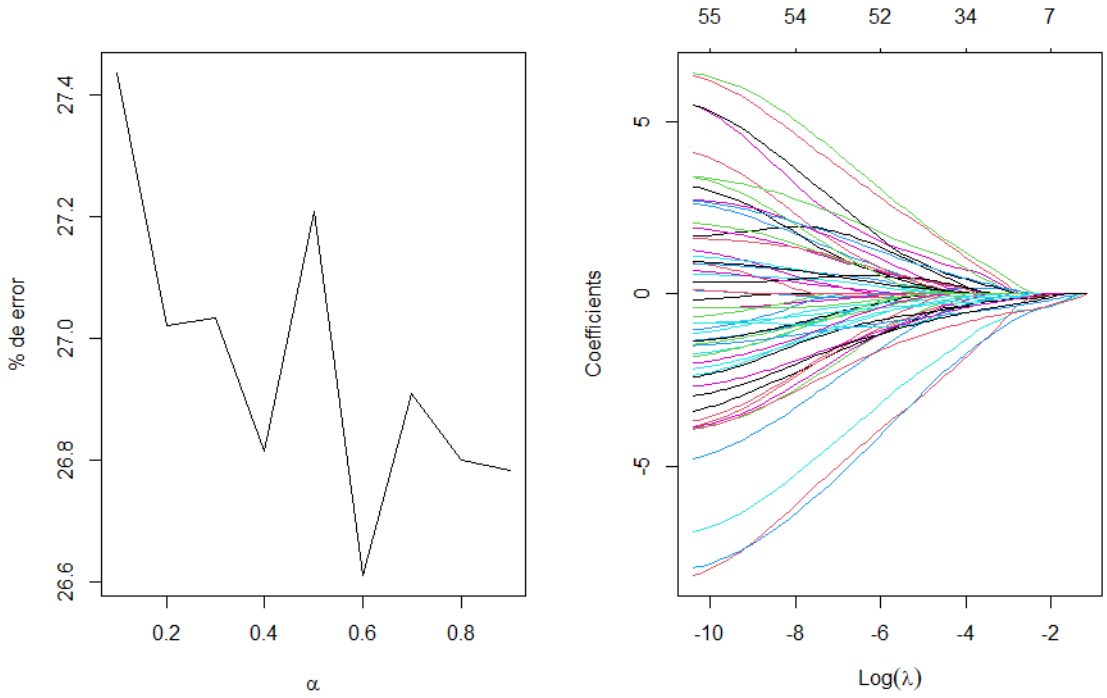

Figura 3.2: Lado izquierdo: Comparación de las tasas de errores (%) para los distintos valores de α.
Lado derecho: Comportamiento del valor de los coeficientes con respecto al valor λ en la milésima repetición.

3.2.4. Análisis Discriminante

Se aplicó LDA y QDA al modelo aditivo, obteniendo unas tasas errores de validación globales de 49.91 % para LDA y 24.77 % en QDA. En la Tabla 3.7 se encuentran los errores de validación por grupo para cada método.

Tabla 3.7: Errores de validación (%) para los métodos de LDA y QDA. Se simularon 200 observaciones para entrenamiento y 200 para validación repitiendo el proceso mil veces.

	Error de Validación		
Modelo	Global	Clase 0	Clase1
LDA	49.91	49.90	49.92
QDA	24.77	24.92	24.61

3.2.5. *Random Forest*

Se aplicó *random forest* al modelo aditivo y se analizaron los parámetros *mtry* y *ntree* con rangos: $mtry \in \{1, 2, \ldots, 10\}$ y $ntree \in \{100, 300, 500, 1000\}$. La Figura 3.3

expone que la mejor configuración para el algoritmo fue *mtry*=2 y *ntree* = 1000. Se obtuvo una tasa de error global de 33.76 % con una desviación estándar de la media del error global de 0.11.

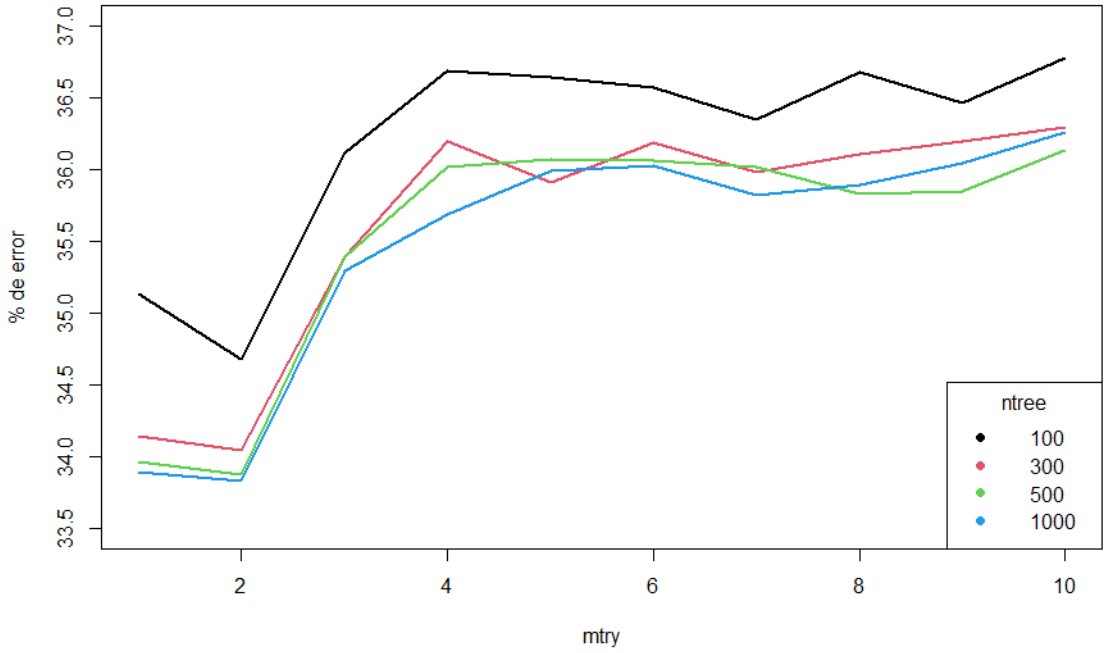

Figura 3.3: Comparación de las tasas de errores con respecto a los valores *mtry* y *ntree*.

3.2.6. Resultados

En la Tabla 3.8 se exponen las tasas de errores de validación para los métodos descritos en este trabajo. Se observa que, en general, los resultados son mejores a los expuestos en la Tabla 3.4. En este caso el método con mejor resultado fue el de QDA.

Nuevamente los métodos de regularización se desempeñaron bien, obteniendo de los mejores resultados entre los métodos. Los métodos de regresión logística a la que se le aplicó *stepwise selection* obtuvieron una tasa de error similar entre los criterios BIC y AIC y el promedio del número coeficientes utilizados en los modelos fue mayor para el criterio AIC con 32.39 coeficientes a comparación de los 20.43 del criterio BIC.

Al simular, se pueden controlar la cantidad de datos. Esto puede mejorar el poder predictivo de los modelos. En este caso se utilizó una base de datos de 200 observaciones para el entrenamiento y una de 200 observaciones para la validación, abriendo la posibilidad de experimentar con una cantidad mayor. La decisión de haber trabajado con 200 observaciones, en el conjunto de entrenamiento, fue para una mejor comparación con las 186 observaciones provenientes del estudio de cáncer de mama.

Si bien no se puede comparar directamente los resultados obtenidos en este tra-

bajo con los reportados en el estudio de simulación hecho por Eslava, G. & Pérez, G. (2022), puesto que en dicho estudio no se utilizó un conjunto de prueba de 200 observaciones, se puede observar una similitud en cuanto al desempeño entre los métodos presentados en ambos trabajos. En ambos casos el método que peor se desempeñó fue la regresión logística, seguida del análisis discriminante lineal, *random forest*, regresión logística con interacciones dos a dos, regresión logística aplicando *stepwise selection* y finalmente análisis discriminante cuadrático. Para LDA se reportó en el estudio de simulación un error de 49.8 % comparado con un error de 49.91 % obtenido en este trabajo, en el caso de la regresión logística aditiva se reportó en el estudio un error de 49.8 % y en este trabajo un error de 50.07 %, para la regresión logística con interacciones dos a dos se reportó en el estudio un error de 26.7 % comparado con el obtenido en este trabajo de 29.46 %.
En ambos casos, como fue mencionado, el método que mejor se desempeñó fue QDA con un error de validación global reportado en el estudio de 24.7 % y un error de 24.77 % reportado en este trabajo.

En la Figura 3.4 se exponen a los errores de validación, por clase y globales, de cada método aplicado al conjunto de datos `DatosSim`.

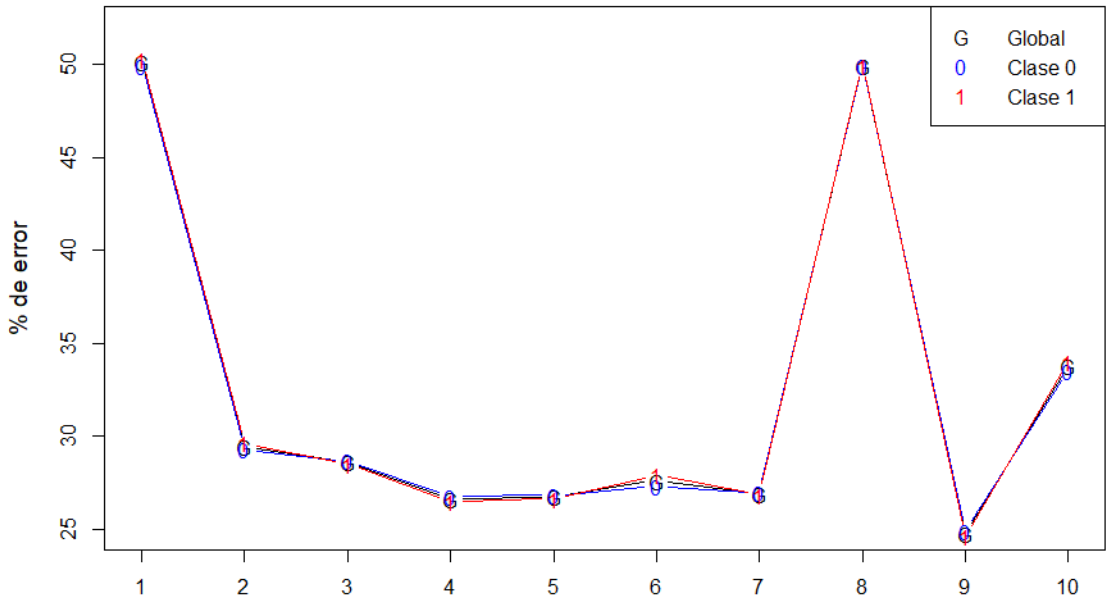

Figura 3.4: Errores de validación (%), por clase y global, al aplicar cada método de este trabajo.
Modelos: 1.Aditivo 2.Interacciones 2 a 2 3.AIC 4.BIC 5.*Ridge Regression* 6.*Lasso* 7.*Elastic Net* 8.LDA 9.QDA 10.*Random Forest*.

Tabla 3.8: Comparación, entre los modelos, de tasas de errores de validación (%) y desviación estándar de la media del error global. Se simularon 200 observaciones para entrenamiento y 200 para validación repitiendo el proceso mil veces.

	Error de Validación			
Modelo	Global	Clase 0	Clase1	SE
Regresión logística aditiva	50.07	49.92	50.23	0.11
Regresión con interacciones 2 a 2	29.46	29.30	29.62	0.11
Regresión con *stepwise selection* AIC	28.59	28.67	28.50	0.12
Regresión con *stepwise selection* BIC	26.63	26.78	26.49	0.12
Regresión regularizada por Ridge	27.61	27.31	27.91	0.11
Regresión regularizada por Lasso	26.74	26.84	26.64	0.11
Regresión regularizada por E. Net	26.89	26.95	26.83	0.11
Análisis discriminante lineal LDA	49.91	49.90	49.92	0.10
Análisis discriminante cuadrático QDA	**24.77**	24.92	24.61	0.10
Random Forest	33.76	33.53	33.99	0.11

Nota: Los modelos de regresión con interacciones 2 a 2, *stepwise selection* y regularizados fueron aplicados al modelo con interacciones 2 a 2. LDA, QDA y *random forest* fueron aplicados al modelo aditivo.

Finalmente, en este capítulo se emplearon los métodos descritos en este trabajo a los conjuntos de datos provenientes de un estudio de cáncer de mama (`DatosCancer`) y un estudio de simulación (`DatosSim`), se reportaron los errores obtenidos y se comparó los resultados obtenidos entre ellos.

Conclusiones

En cuanto al poder predictivo de los métodos presentados en esta tesis, para el conjunto de datos del estudio de cáncer de mama, los mejores resultados se obtuvieron con el modelo de regresión logística a la que se le aplicó *backward stepwise selection, lasso* y *elastic net*. Los errores de validación estuvieron entre 33-34 %, si bien *backward stepwise selection* es el único método algorítmico de los tres, *lasso* y *elastic net* se pueden representar como problemas de optimización que son resueltos utilizando algoritmos. Estos tres métodos se destacan por utilizar un poder computacional considerable para su aplicación. Por otro lado, *random forest* obtuvo un error de validación de 37 % y se tardó 15 minutos en calcular los errores. Se puede concluir que el método con mayor poder, en general, fue la clásica regresión logística a la que se le aplicó el método de selección de variables *stepwise selection* con el criterio BIC con un error de validación de 33.25 %. Esto puede sorprender porque al existir métodos más modernos (*random forest*), el que mejor se desempeñó fue la regresión logística que es un método clásico. Todos los métodos arrojaron una desviación estándar de la media del error global similar de entre 0.22-0.34.

Para el caso de los datos simulados provenientes de una distribución Gaussiana Condicional, el método de mejor desempeño fue QDA con un error de 24.77 %, si bien se simularon los datos con matriz de covarianzas distintas para cada clase, no se debe de olvidar que los datos provienen de una combinación de variables discretas y continuas, por lo que no se puede dar una razón genuina del desempeño del método QDA en los datos, puesto que es un método exclusivo para variables continuas. En general, varios métodos son exclusivos para el caso de variables continuas aunque normalmente se desempeñan de buena manera para el caso de variables mixtas. En este conjunto de datos se obtuvieron errores de validación menores que en el caso del estudio de cáncer de mama, si bien la mayoría de los errores se encuentran entre 24-28 %, hay métodos en el rango de 50 % de error como la regresión logística aditiva y LDA. En este caso, la desviación estándar para la media de los errores de validación globales fue menor que en el estudio de cáncer de mama con un rango entre 0.10-0.12. En ambos conjuntos utilizando a la regresión logística con *stepwise selection* y criterio BIC se obtuvieron de los mejores desempeños.

Al aplicar los métodos expuestos en este trabajo, pude observar que hay métodos que son más directos de aplicar como lo son la regresión logística, la regresión logística con *backward stepwise selection*, los métodos de regularización: *ridge regression, lasso*, LDA y QDA y hay métodos como *random forest* que necesitan una análisis previo.

Por otro lado, hay diferencias en la demanda del poder computacional entre cada método: hay métodos que necesitan un poder computacional estándar para su aplicación y tardan 47 segundos en calcular los errores como fue mi caso para la regresión logística aditiva, no obstante, hay métodos que necesitan de un poder computacional considerable y que tardan hasta 2 horas en calcular los errores de validación como fue el caso de *elastic net*. De los métodos empleados en este trabajo, los métodos de regularización (*lasso, ridge regression* y *elastic net*) fueron los más demandantes en términos de tiempo tardando hasta dos horas en su implementación, utilizando una PC con procesador Ryzen 7 6800HS y 16GB en memoria RAM.

Los resultados presentados en esta tesis no son definitivos y siempre hay un margen de mejora, exploré las opciones disponibles en su momento, en el caso de *random forest* determiné el número de variables a considerar con una malla desde 1 hasta 9 variables probando de uno en uno, al igual que en *elastic net* para el valor de α el cual lo determiné con una malla desde 0.1 a 0.9 de 0.1 en 0.1. Como Taylor menciona acerca de la comparación entre métodos de clasificación en la discusión de Ripley (1994, p.441):

> La comparación de métodos ... normalmente es difícil de interpretar. Las diferencias observadas de la bondad de los resultados pueden provenir de:
>
> - diferentes adecuaciones de los métodos básicos para los conjuntos de datos dados,
> - diferente sofisticación de procedimientos por defecto para la configuración de parámetros,
> - diferente sofisticación del programador en la selección de opciones y ajuste de los parámetros y
> - la eficacia del procesamiento de datos por parte del usuario.

Puede que un análisis más preciso obtenga mejores resultados, no obstante, hice lo mejor que pude.

Este trabajo fue desafiante, en especial la parte de reportar y escribir. Me di cuenta que los *papers* y libros de ciencia conllevan una extensa revisión y una continua edición para su mayor aprovechamiento dentro de la comunidad. Espero haber cumplido con la tarea de exponer y explicar los métodos de la manera más simple y correcta sin dar cabida a ambigüedades.

Anexo A

Tablas y figuras suplementarias

Tabla A.1: Eigenvalores calculados con las variables estandarizadas con media cero y varianza uno y sin estandarizar del conjunto de datos `DatosCancer` y el porcentaje de varianza explicada por cada uno.

	Estandarizadas	% de varianza	Sin estandarizar	% de varianza
PC1	2.29	25.46	2,002,778.29	92.67
PC2	1.75	19.52	112,332.03	5.19
PC3	1.17	13.07	45,173.31	2.09
PC4	0.99	11.07	802.75	0.00
PC5	0.87	9.75	15.64	0.00
PC6	0.76	8.44	1.77	0.00
PC7	0.53	5.96	0.31	0.00
PC8	0.37	4.13	0.25	0.00
PC9	0.23	2.59	0.13	0.00

Tabla A.2: Eigenvectores calculados con las variables estandarizadas con media cero y varianza uno del conjunto de datos `DatosCancer`.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
x_1	0.1897	0.3784	-0.2043	0.3979	-0.2153	0.7283	-0.0795	-0.1576	-0.1133
x_2	0.0670	-0.5962	-0.3018	0.0844	0.1095	0.3372	-0.0461	0.6404	0.0628
x_3	-0.3807	-0.0192	0.2262	-0.4724	-0.3389	0.4182	0.5224	0.0543	0.1217
x_4	-0.3647	0.0491	-0.3137	0.5653	-0.0998	-0.2943	0.5729	0.0866	-0.1200
x_5	-0.5828	-0.0653	0.0873	-0.0263	-0.0147	0.0784	-0.4050	0.0445	-0.6895
x_6	-0.5747	0.0630	-0.0971	0.1632	0.0851	0.0704	-0.3772	-0.0943	0.6829
x_7	0.0039	-0.6589	0.0142	0.1416	0.0892	0.1501	0.1142	-0.7076	-0.0355
x_8	-0.0396	0.0805	0.6597	0.3291	0.5902	0.2027	0.1896	0.1520	0.0058
x_9	-0.1008	0.2230	-0.5150	-0.3690	0.6731	0.1403	0.1875	-0.1411	-0.1046

Tabla A.3: Eigenvectores calculados con las variables sin estandarizar del conjunto de datos `DatosCancer`.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
x_1	-0.0268	-0.0286	0.0055	-0.9865	-0.1442	0.0604	0.0228	0.0150	0.0004
x_2	0.0018	0.0013	0.0005	-0.0575	0.0338	-0.9312	0.1796	0.0242	0.3090
x_3	0.0063	0.0013	0.0004	-0.1405	0.9884	0.0525	0.0189	0.0106	0.0120
x_4	-0.1736	-0.3146	0.9330	0.0188	0.0015	0.0001	0.0002	0.0001	0.0001
x_5	-0.7358	0.6713	0.0894	0.0016	0.0036	0.0003	0.0001	0.0001	0.0000
x_6	-0.6540	-0.6705	-0.3485	0.0352	0.0002	0.0002	0.0001	0.0003	0.0000
x_7	0.0010	0.0006	0.0003	-0.0295	0.0288	-0.3422	-0.3075	0.1553	-0.8733
x_8	0.0009	0.0006	0.0000	-0.0263	0.0123	-0.0574	-0.9305	0.0395	0.3584
x_9	0.0008	0.0008	0.0000	-0.0236	0.0143	-0.0775	-0.0807	-0.9866	-0.1154

Tabla A.4: Eigenvalores calculados con las variables estandarizadas con media cero y varianza uno y sin estandarizar del conjunto de datos `DatosSim` y el porcentaje de varianza explicada por cada uno.

	Estandarizadas	% de varianza	Sin estandarizar	% de varianza
PC1	1.570	13.30	9.455	60.04
PC2	1.357	12.76	1.400	7.11
PC3	1.243	12.22	1.152	6.95
PC4	1.020	11.03	0.935	6.94
PC5	0.950	10.41	0.851	5.57
PC6	0.882	9.27	0.778	4.97
PC7	0.846	8.76	0.657	4.61
PC8	0.750	7.90	0.240	1.65
PC9	0.713	7.32	0.226	1.31
PC10	0.670	6.98	0.214	1.25

Tabla A.5: Eigenvectores calculados con las variables estandarizadas con media cero y varianza uno del conjunto de datos `DatosSim`.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
i_1	0.0284	-0.0269	0.6790	-0.1687	0.3133	-0.1343	0.2790	0.2406	0.3405	0.3757
i_2	-0.2833	-0.4520	0.0055	0.1113	0.3263	-0.4560	-0.3647	0.4167	-0.2217	-0.1813
i_3	-0.3785	0.0428	0.2919	-0.4074	-0.3360	0.4327	-0.2761	0.3097	0.1168	-0.3432
i_4	0.2542	-0.2970	0.2904	-0.2104	-0.5310	-0.4297	-0.3216	-0.3748	0.0244	0.0862
x_1	0.4763	0.1652	0.1898	-0.3826	0.2015	0.0859	-0.0147	0.1134	-0.7013	-0.1024
x_2	-0.4768	-0.3422	-0.0495	-0.1537	-0.1536	0.1827	0.2560	-0.1239	-0.4587	0.5304
x_3	0.0889	-0.5661	0.1009	-0.0862	0.4068	0.3716	0.0453	-0.4725	0.1342	-0.3274
x_4	0.2272	-0.1302	-0.5065	-0.5126	0.1628	0.0973	-0.3085	0.1851	0.3039	0.3930
x_5	0.3244	-0.2158	0.1856	0.5544	-0.1321	0.4609	-0.3462	0.2453	-0.0617	0.3043
x_6	-0.3035	0.4203	0.1724	0.0369	0.3562	0.0356	-0.5716	-0.4315	-0.0659	0.2341

Tabla A.6: Eigenvectores calculados con las variables sin estandarizar del conjunto de datos `DatosSim`.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
i_1	-0.4959	0.0346	-0.0507	0.0222	-0.0540	0.0070	-0.0490	-0.2118	-0.2134	-0.8088
i_2	-0.5030	-0.0881	0.0661	-0.0589	0.0216	-0.0471	0.0154	0.7517	-0.3552	0.1930
i_3	-0.4935	-0.0888	-0.0540	-0.0860	-0.0279	0.0099	0.0636	-0.6108	-0.2789	0.5315
i_4	-0.5012	0.0728	0.0379	-0.0120	0.0694	0.0273	0.0104	0.0557	0.8539	0.0628
x_1	-0.0208	0.7585	-0.1588	-0.1080	-0.5108	0.3119	0.1265	0.0679	-0.0515	0.0771
x_2	0.0377	-0.4870	0.3385	-0.1831	-0.4399	0.4074	0.4914	0.0070	0.0687	-0.0870
x_3	-0.0249	0.0720	0.6418	0.1913	-0.4782	-0.4107	-0.3772	-0.0612	0.0105	0.0479
x_4	0.0480	0.2614	0.3233	-0.7121	0.2654	-0.3782	0.3084	-0.0495	-0.0227	-0.0789
x_5	-0.0357	0.2204	0.1861	0.6287	0.1842	-0.1991	0.6657	-0.0231	-0.0546	0.0030
x_6	0.0156	-0.2081	-0.5458	-0.0662	-0.4512	-0.6183	0.2349	0.0503	0.1016	-0.0156

Tabla A.7: Correlaciones redondeadas a dos decimales entre las variables explicativas sin estandarizar $\mathtt{x}_1, \ldots, \mathtt{x}_9$ y los componentes principales para el conjunto de datos `DatosCancer`.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	$\mathtt{x}_1$	$\mathtt{x}_2$	$\mathtt{x}_3$	$\mathtt{x}_4$	$\mathtt{x}_5$	$\mathtt{x}_6$	$\mathtt{x}_7$	$\mathtt{x}_8$	$\mathtt{x}_9$
PC1	1.00	-0.23	-0.03	0.88	0.07	-0.13	-0.09	-0.05	-0.04	0.19	0.05	-0.41	-0.44	-0.95	-0.90	-0.02	-0.06	-0.10
PC2	-0.23	1.00	0.00	0.10	0.01	-0.01	-0.01	-0.01	0.00	-0.24	0.06	0.27	-0.29	0.53	-0.19	0.12	0.05	-0.13
PC3	-0.03	0.00	1.00	0.01	0.00	0.00	0.00	0.00	0.00	-0.01	-0.03	0.07	-0.78	0.00	0.17	-0.04	0.08	0.08
PC4	0.88	0.10	0.01	1.00	-0.03	0.05	0.04	0.02	0.02	0.44	0.02	-0.30	-0.55	-0.74	-0.92	-0.07	-0.04	-0.13
PC5	0.07	0.01	0.00	-0.03	1.00	0.00	0.00	0.00	0.00	-0.33	-0.02	0.82	-0.04	-0.06	-0.07	0.07	-0.05	-0.03
PC6	-0.13	-0.01	0.00	0.05	0.00	1.00	-0.01	0.00	0.00	0.48	-0.90	0.22	0.08	0.11	0.13	-0.66	0.10	0.04
PC7	-0.09	-0.01	0.00	0.04	0.00	-0.01	1.00	0.00	0.00	0.33	0.27	0.14	0.06	0.07	0.09	-0.15	-0.87	0.08
PC8	-0.05	-0.01	0.00	0.02	0.00	0.00	0.00	1.00	0.00	0.17	0.10	0.07	0.03	0.04	0.05	0.27	0.13	-0.93
PC9	-0.04	0.00	0.00	0.02	0.00	0.00	0.00	0.00	1.00	0.16	0.19	0.07	0.03	0.04	0.04	-0.57	0.35	0.00
$\mathtt{x}_1$	0.19	-0.24	-0.01	0.44	-0.33	0.48	0.33	0.17	0.16	1.00	-0.13	-0.15	0.00	-0.25	-0.09	-0.28	-0.01	0.04
$\mathtt{x}_2$	0.05	0.06	-0.03	0.02	-0.02	-0.90	0.27	0.10	0.19	-0.13	1.00	-0.08	-0.03	-0.03	-0.08	0.57	-0.16	-0.04
$\mathtt{x}_3$	-0.41	0.27	0.07	-0.30	0.82	0.22	0.14	0.07	0.07	-0.15	-0.08	1.00	0.06	0.44	0.31	0.00	0.00	0.01
$\mathtt{x}_4$	-0.44	-0.29	-0.78	-0.55	-0.04	0.08	0.06	0.03	0.03	0.00	-0.03	0.06	1.00	0.31	0.45	-0.01	-0.05	0.05
$\mathtt{x}_5$	-0.95	0.53	0.00	-0.74	-0.06	0.11	0.07	0.04	0.04	-0.25	-0.03	0.44	0.31	1.00	0.72	0.06	0.07	0.04
$\mathtt{x}_6$	-0.90	-0.19	0.17	-0.92	-0.07	0.13	0.09	0.05	0.04	-0.09	-0.08	0.31	0.45	0.72	1.00	-0.04	0.05	0.16
$\mathtt{x}_7$	-0.02	0.12	-0.04	-0.07	0.07	-0.66	-0.15	0.27	-0.57	-0.28	0.57	0.00	-0.01	0.06	-0.04	1.00	0.01	-0.20
$\mathtt{x}_8$	-0.06	0.05	0.08	-0.04	-0.05	0.10	-0.87	0.13	0.35	-0.01	-0.16	0.00	-0.05	0.07	0.05	0.01	1.00	-0.10
$\mathtt{x}_9$	-0.10	-0.13	0.08	-0.13	-0.03	0.04	0.08	-0.93	0.00	0.04	-0.04	0.01	0.05	0.04	0.16	-0.20	-0.10	1.00

Tabla A.8: Correlaciones redondeadas a dos decimales entre las variables explicativas sin estandarizar y los componentes principales del conjunto de datos `DatosSim`.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	i_1	i_2	i_3	i_4	x_1	x_2	x_3	x_4	x_5	x_6
PC1	1.00	-0.08	0.08	0.06	-0.03	0.10	0.02	-0.01	-0.07	0.08	-0.55	-0.59	-0.48	-0.52	-0.14	-0.04	0.09	-0.07	-0.01	0.04
PC2	-0.08	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.08	-0.05	0.07	0.11	0.45	-0.43	-0.29	0.49	-0.58	-0.37
PC3	0.08	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.03	-0.04	-0.12	0.01	-0.43	-0.32	-0.28	-0.02	-0.52	0.69
PC4	0.06	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	-0.05	0.12	-0.11	-0.52	-0.45	-0.21	-0.46	0.00	-0.56
PC5	-0.03	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	-0.11	0.04	0.10	-0.06	0.57	-0.35	-0.23	-0.58	0.18	0.23
PC6	0.10	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.08	-0.17	0.05	-0.22	0.02	-0.47	0.82	0.03	-0.03	0.06
PC7	0.02	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.12	-0.11	-0.02	0.11	0.38	0.25	-0.45	-0.60	-0.11
PC8	-0.01	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.33	-0.29	0.66	-0.63	0.01	0.07	-0.05	0.02	-0.03	0.02
PC9	-0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.77	-0.10	-0.48	-0.05	0.03	-0.02	-0.04	-0.01	0.01	-0.01
PC10	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	-0.75	0.12	0.48	0.00	0.03	0.01	-0.04	-0.01	0.00
i_1	-0.55	-0.08	-0.03	0.03	-0.11	0.08	0.00	0.33	0.77	0.00	1.00	0.09	0.11	0.01	-0.06	0.03	0.07	0.02	0.03	-0.02
i_2	-0.59	-0.05	-0.04	-0.05	0.04	-0.17	0.12	-0.29	-0.10	-0.75	0.09	1.00	0.02	0.15	0.05	0.13	-0.07	-0.05	0.01	0.00
i_3	-0.48	0.07	-0.12	0.12	0.10	0.05	-0.11	0.66	-0.48	0.12	0.11	0.02	1.00	-0.08	0.07	-0.10	-0.04	-0.01	0.07	-0.12
i_4	-0.52	0.11	0.01	-0.11	-0.06	-0.22	-0.02	-0.63	-0.05	0.48	0.01	0.15	-0.08	1.00	0.07	0.08	-0.14	0.12	-0.04	-0.01
x_1	-0.14	0.45	-0.43	-0.52	0.57	0.02	0.11	0.01	0.03	0.00	-0.06	0.05	0.07	0.07	1.00	0.01	0.01	0.10	0.00	-0.05
x_2	-0.04	-0.43	-0.32	-0.45	-0.35	-0.47	0.38	0.07	-0.02	0.03	0.03	0.13	-0.10	0.08	0.01	1.00	0.09	0.02	0.14	0.04
x_3	0.09	-0.29	-0.28	-0.21	-0.23	0.82	0.25	-0.05	-0.04	0.01	0.07	-0.07	-0.04	-0.14	0.01	0.09	1.00	0.01	0.10	0.00
x_4	-0.07	0.49	-0.02	-0.46	-0.58	0.03	-0.45	0.02	-0.01	-0.04	0.02	-0.05	-0.01	0.12	0.10	0.02	0.01	1.00	-0.11	-0.02
x_5	-0.01	-0.58	-0.52	0.00	0.18	-0.03	-0.60	-0.03	0.01	-0.01	0.03	0.01	0.07	-0.04	0.00	0.14	0.10	-0.11	1.00	-0.04
x_6	0.04	-0.37	0.69	-0.56	0.23	0.06	-0.11	0.02	-0.01	0.00	-0.02	0.00	-0.12	-0.01	-0.05	0.04	0.00	-0.02	-0.04	1.00

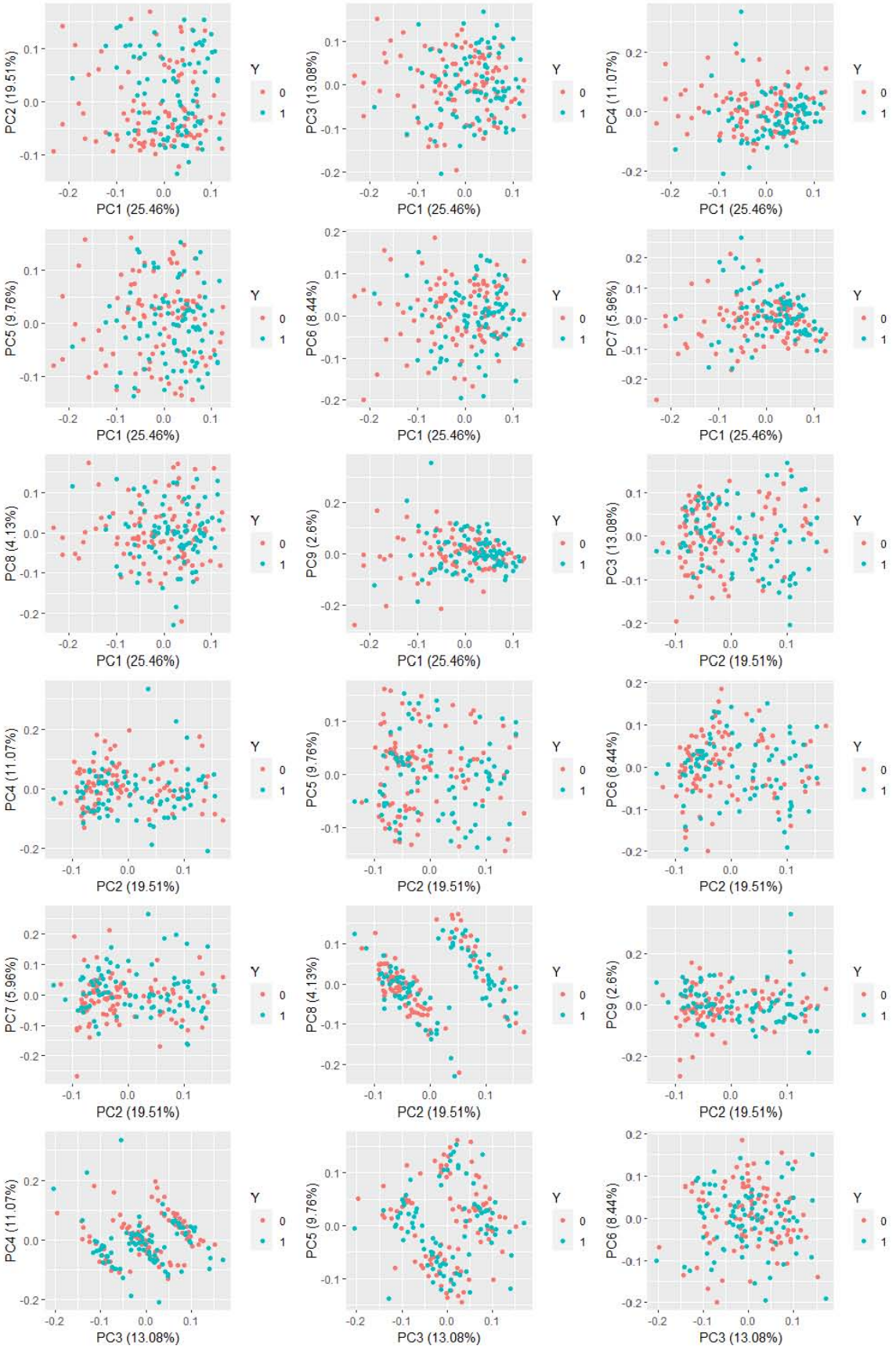

Figura A.1: Componentes principales calculados con las variables estandarizadas con media cero y varianza uno del conjunto de datos `DatosCancer`.

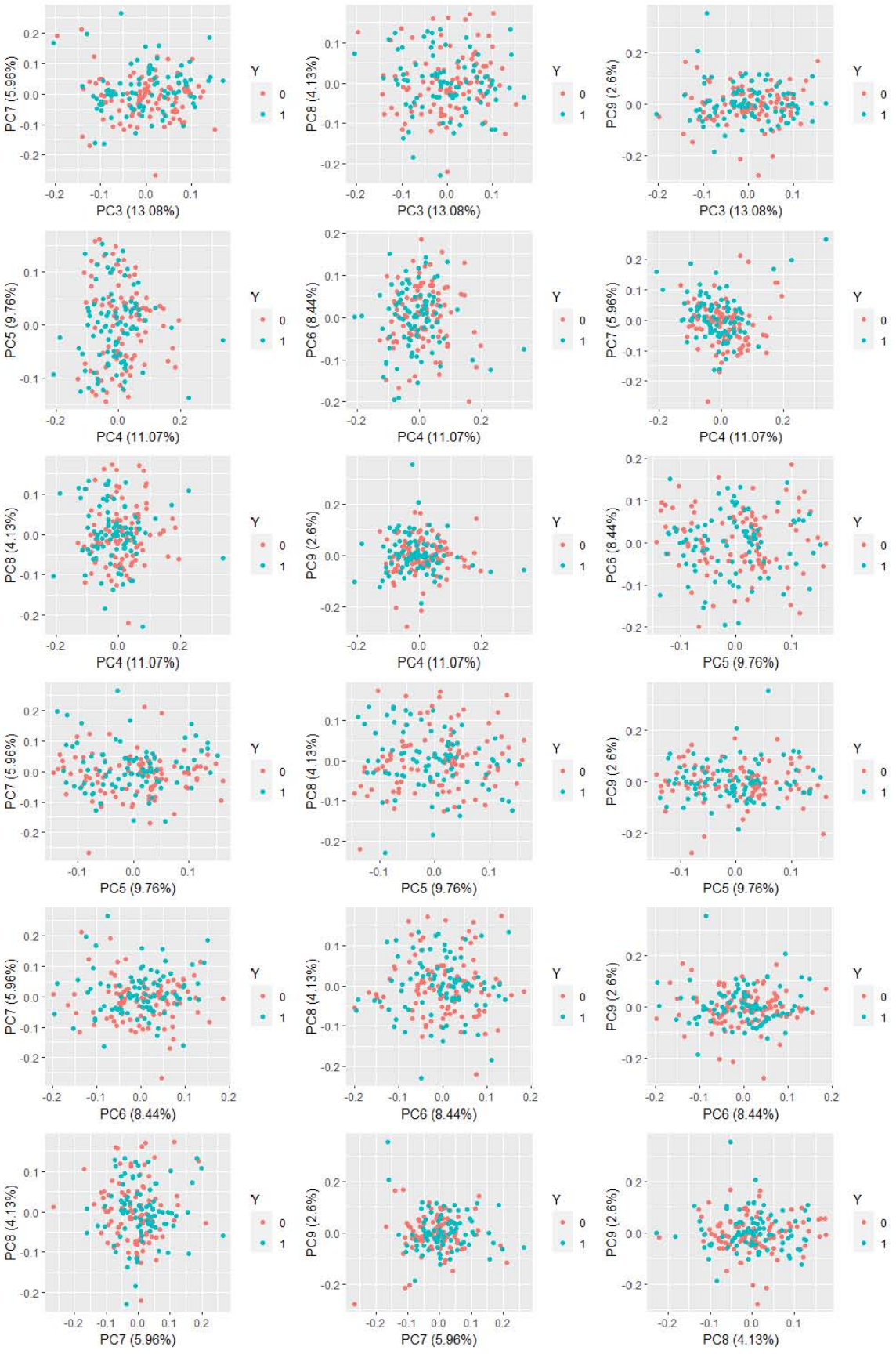

Figura A.2: Continuación de los componentes principales calculados con las variables estandarizadas con media cero y varianza uno del conjunto de datos `DatosCancer`.

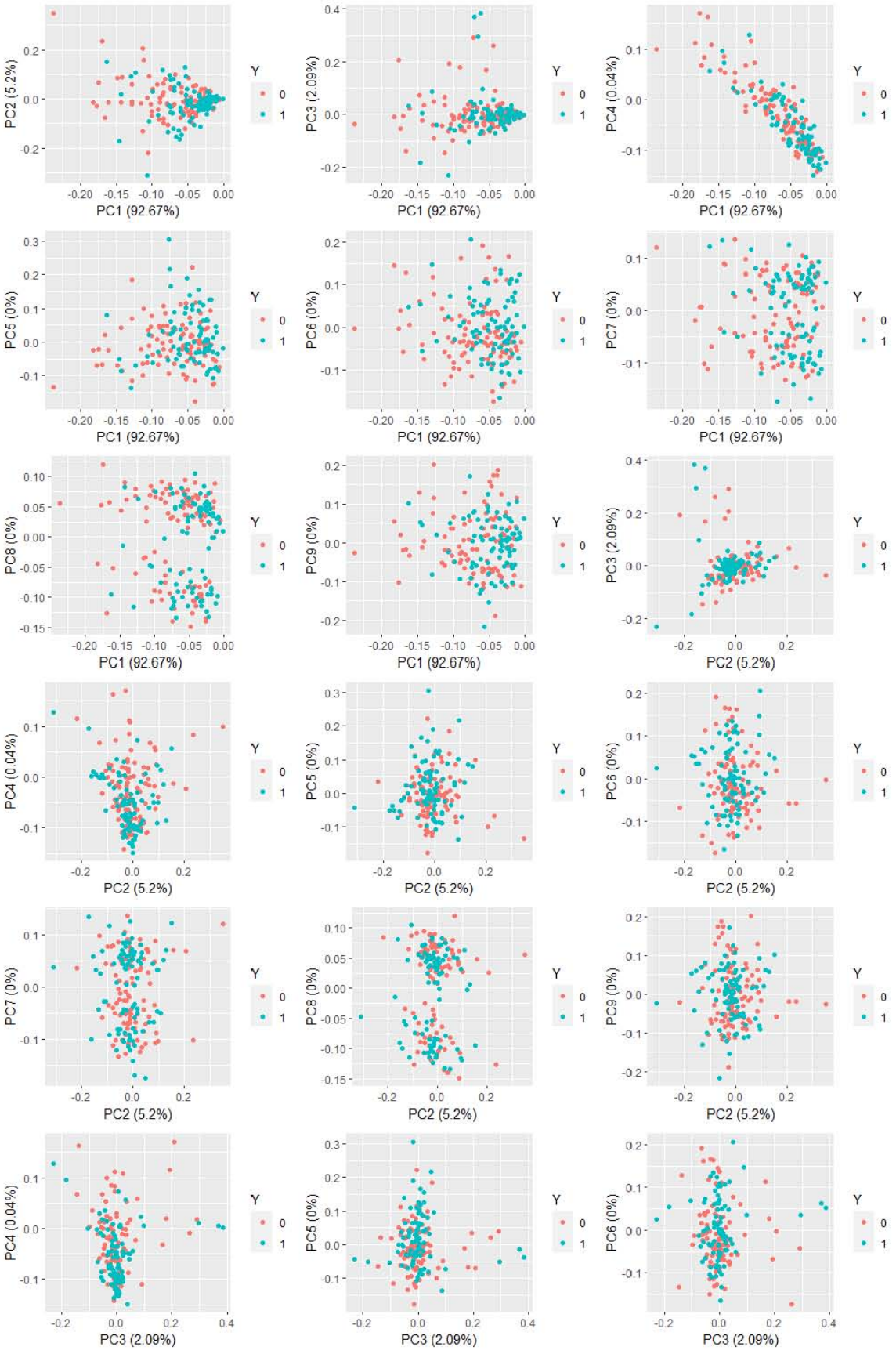

Figura A.3: Componentes principales calculados con las variables sin estandarizar del conjunto de datos `DatosCancer`.

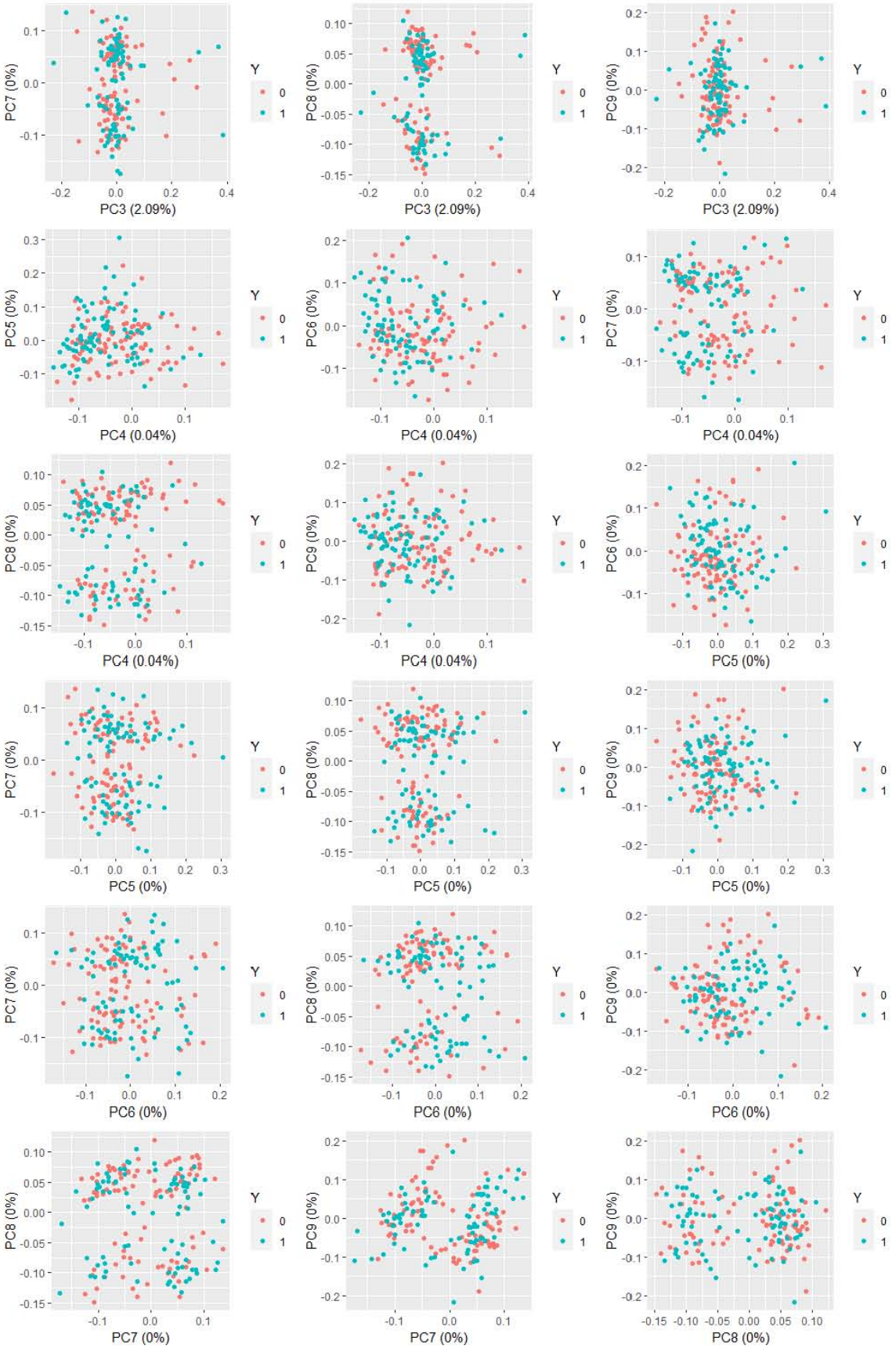

Figura A.4: Continuación de los componentes principales calculados con las variables sin estandarizar del conjunto de datos `DatosCancer`.

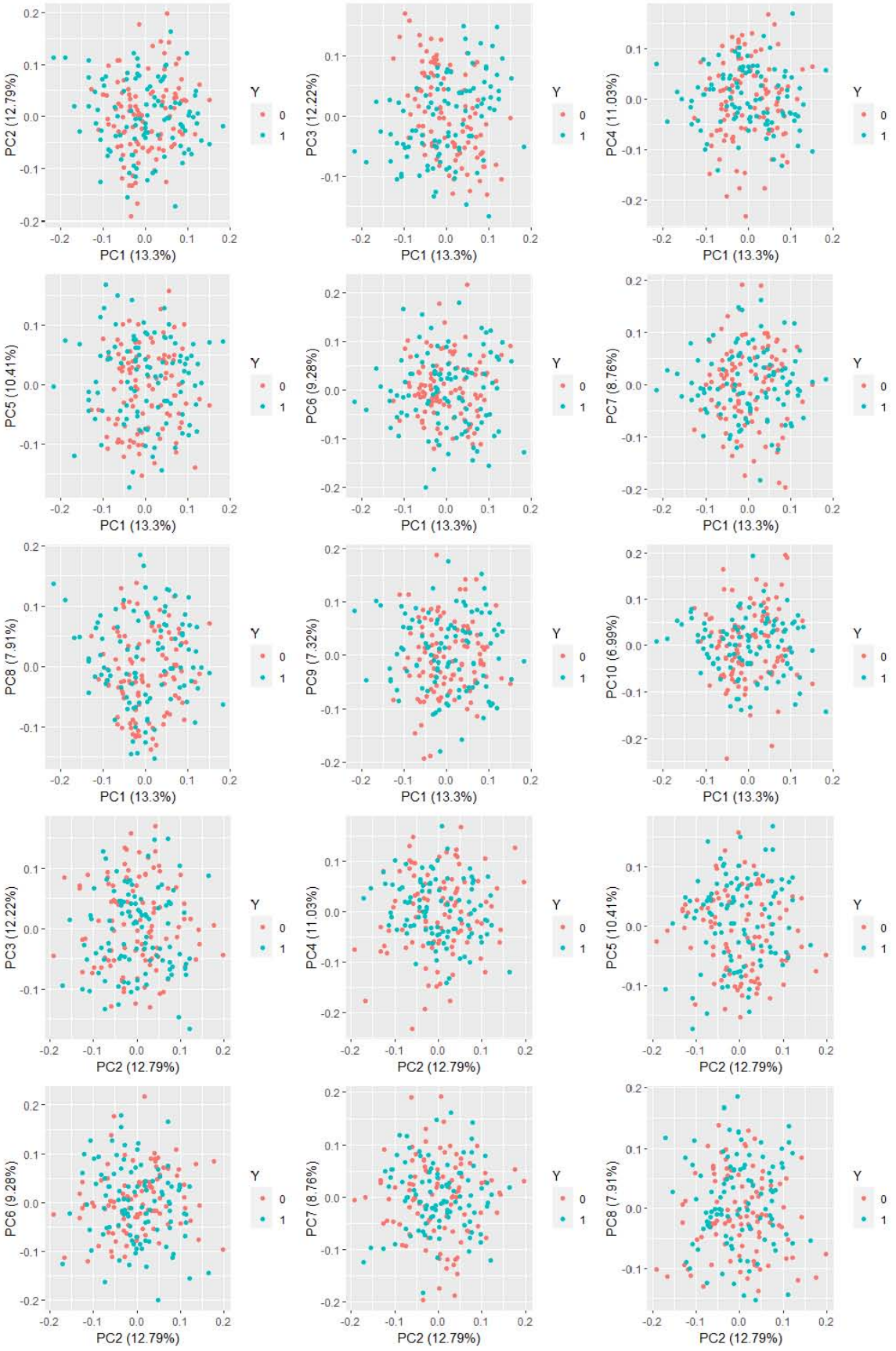

Figura A.5: Componentes principales calculados con las variables estandarizadas con media cero y varianza uno del conjunto de datos `DatosSim`.

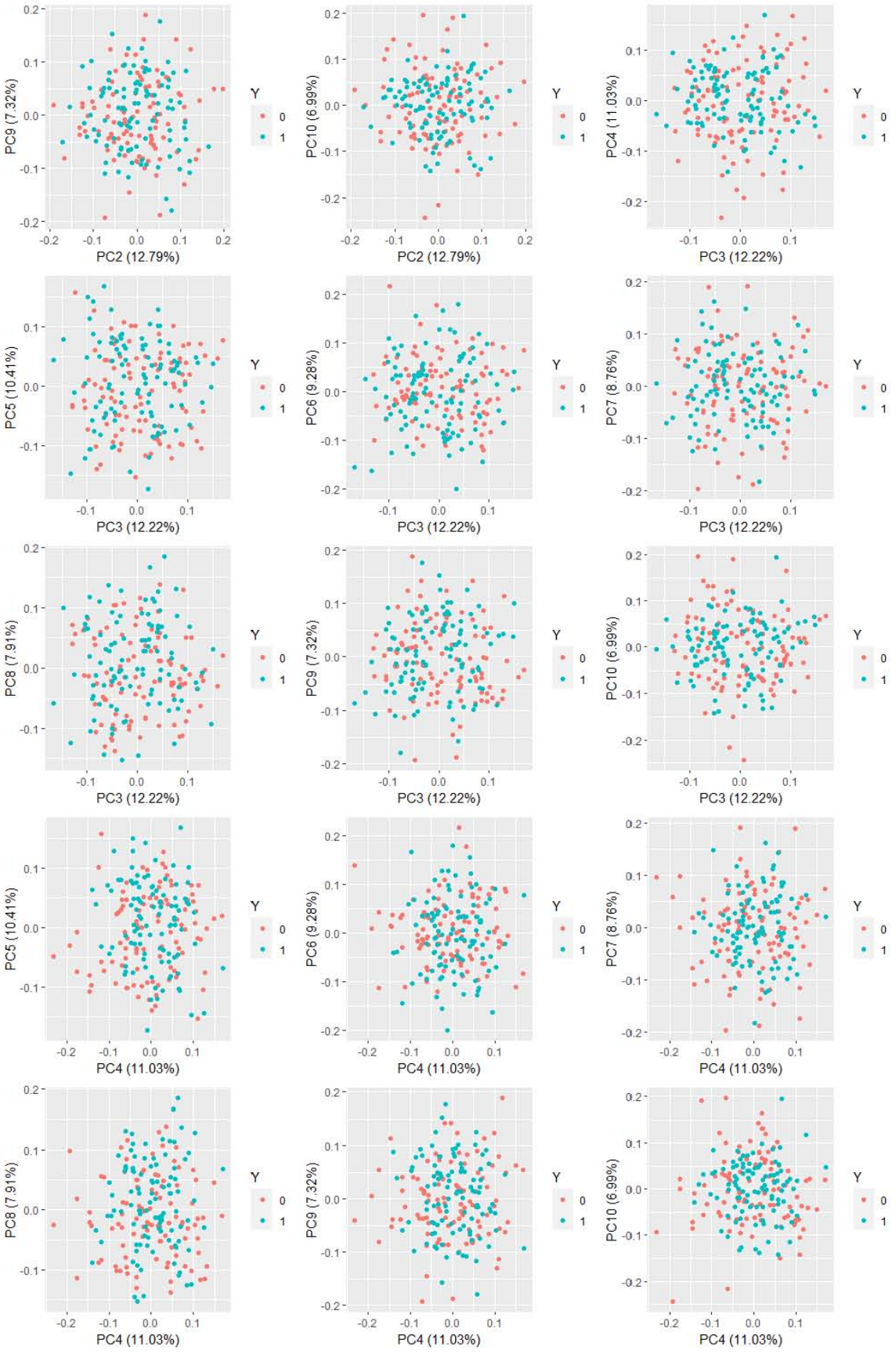

Figura A.6: Continuación de los componentes principales calculados con las variables estandarizadas con media cero y varianza uno del conjunto de datos `DatosSim`.

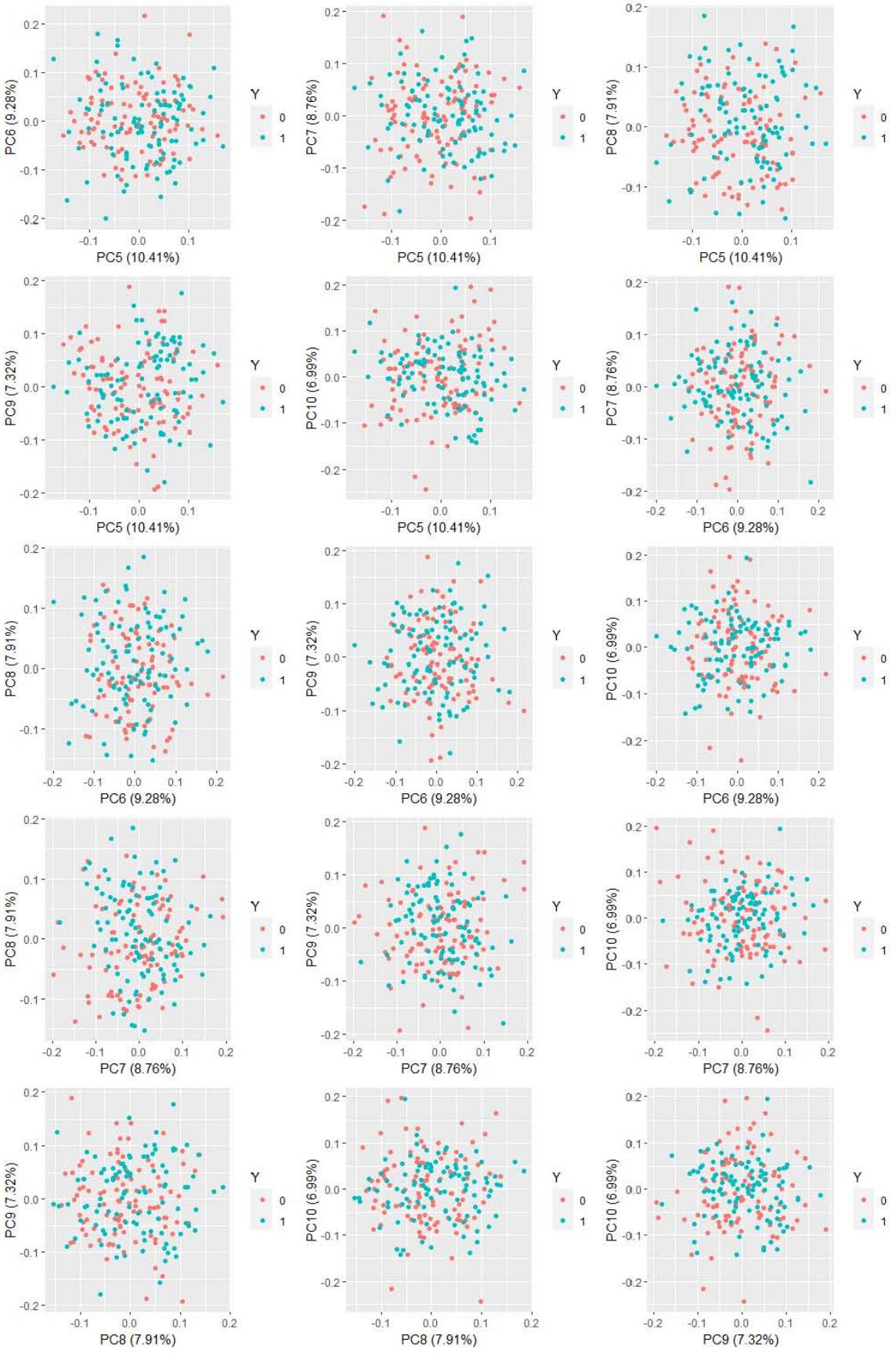

Figura A.7: Continuación de los componentes principales calculados con las variables estandarizadas con media cero y varianza uno del conjunto de datos `DatosSim`.

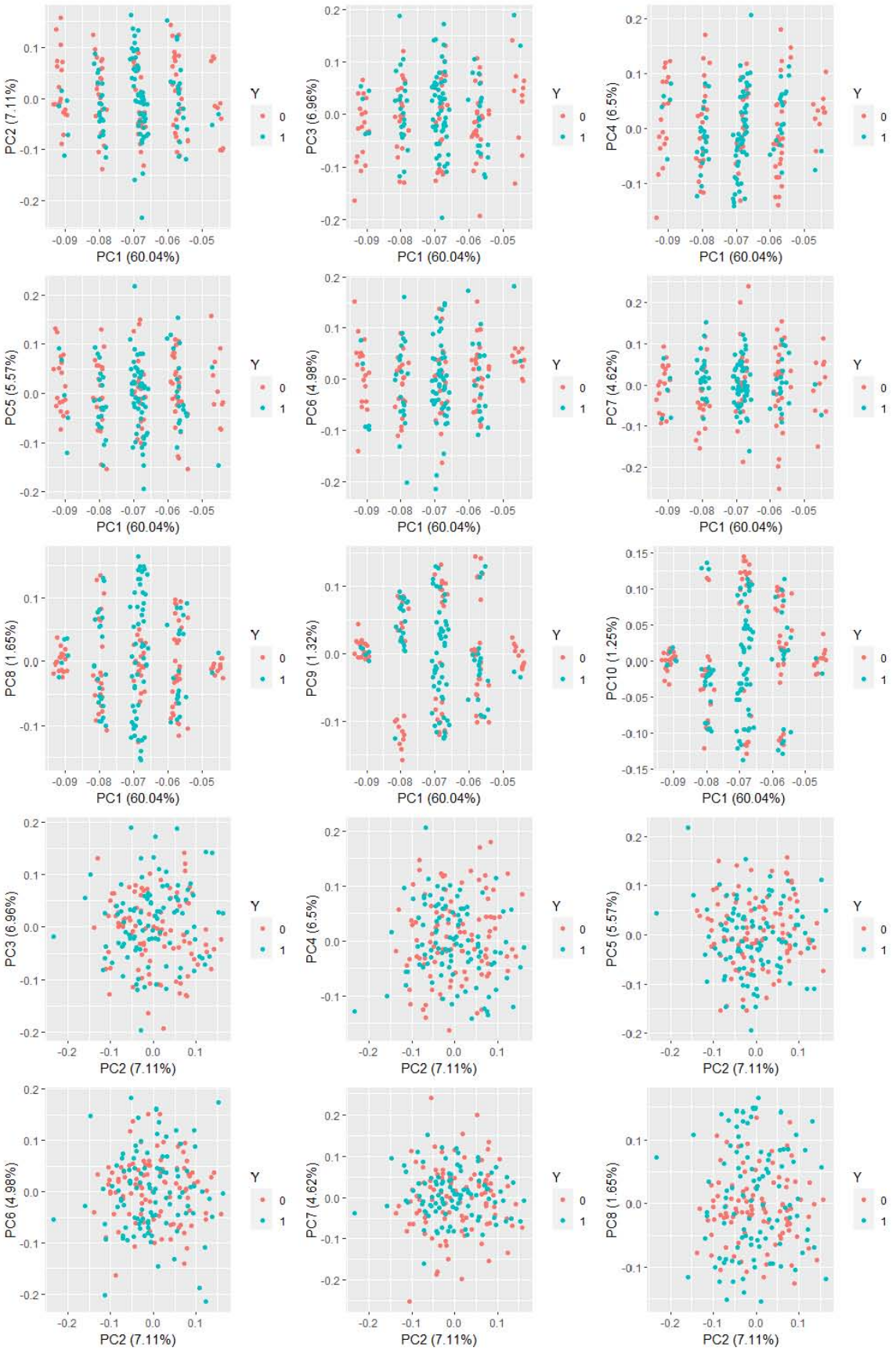

Figura A.8: Componentes principales calculados con las variables sin estandarizar del conjunto de datos `DatosSim`.

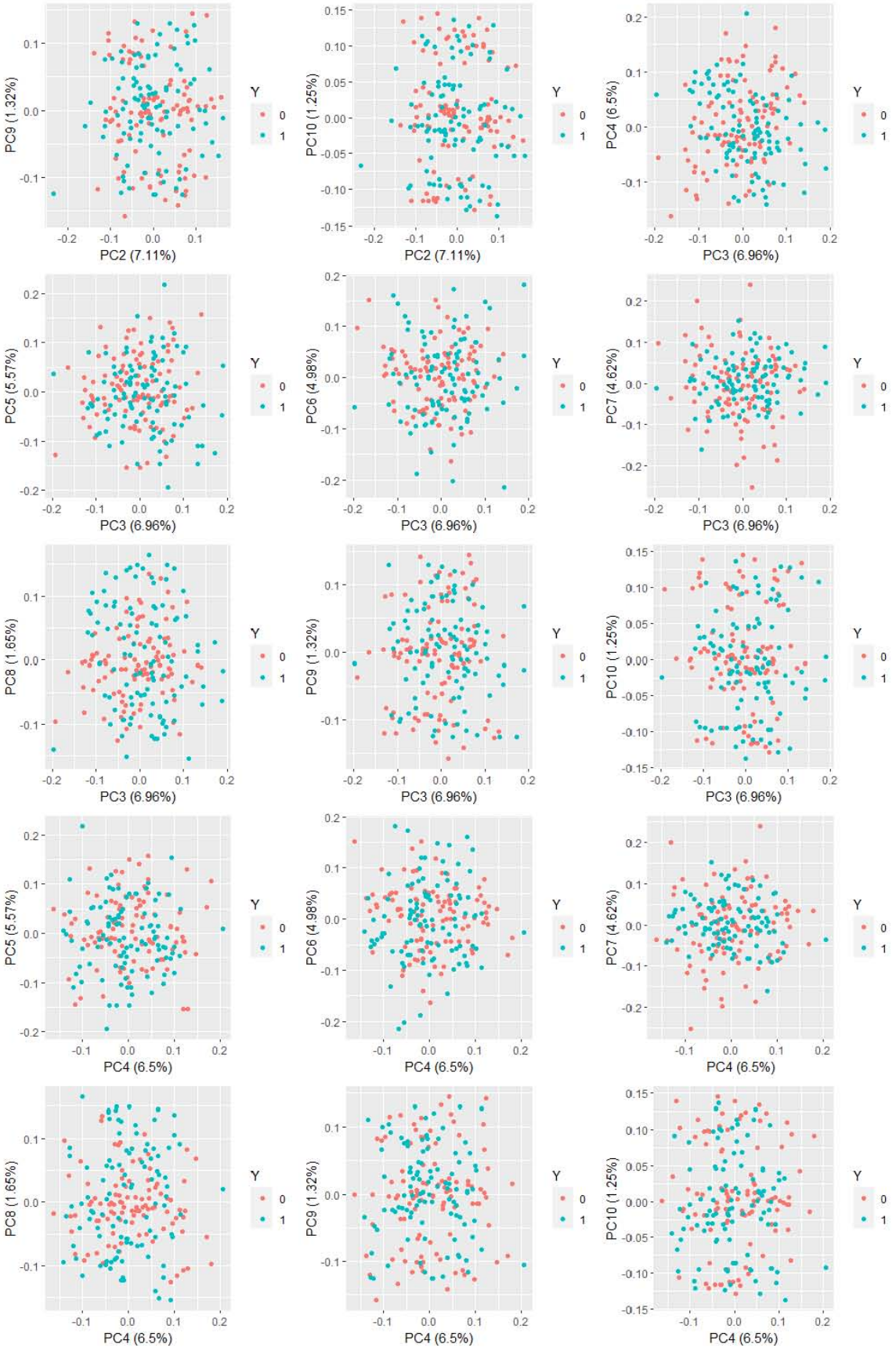

Figura A.9: Continuación de los componentes principales calculados con las variables sin estandarizar del conjunto de datos `DatosSim`.

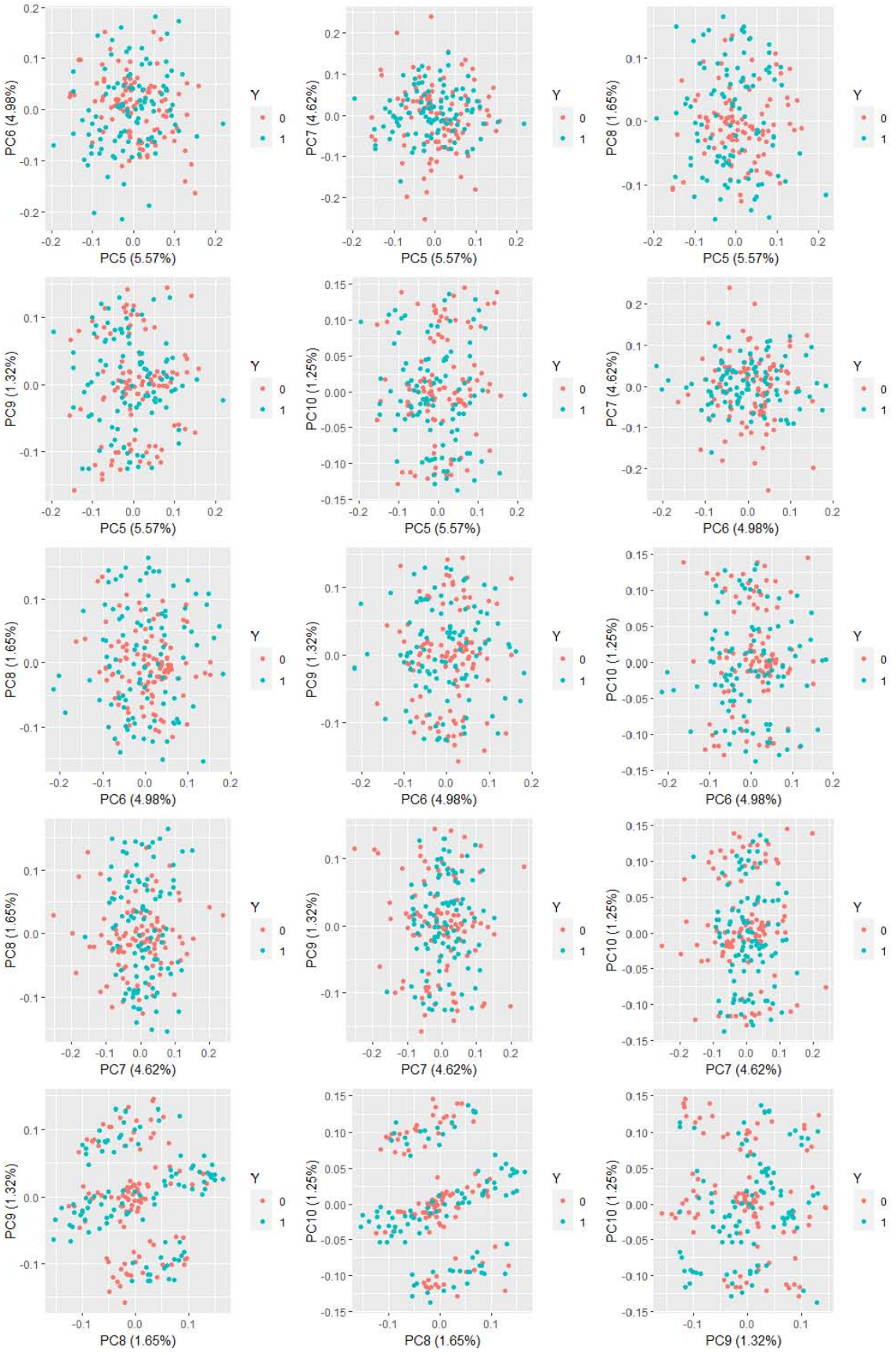

Figura A.10: Continuación de los componentes principales calculados con las variables sin estandarizar del conjunto de datos `DatosSim`.

Anexo B

Paquetería utilizada en **R**

Para la aplicación de los métodos en **R** se utilizaron varias paqueterías que se presentarán a continuación:

- `stats`
  Esta paquetería viene de manera predeterminada con **R** y se le atribuye a R Core Team (2022). En ella se utilizó la función `glm()` para ajustar modelos lineales generalizados como la *Regresión Logística*.
  Para más información acerca de la paquetería, así como otras funciones dentro de ella, se le recomienda al lector visitar
  https://stat.ethz.ch/R-manual/R-devel/library/stats/html/00Index.html
- `glmnet`
  Esta paquetería se le atribuye a Jerome Friedman, Trevor Hastie & Robert Tibshirani (2010). Esta paquetería se encarga de ajustar, de manera eficiente, a los métodos de regularización como *Lasso, Elastic Net* y *Ridge Regression* para los modelos de *Regresión Lineal, Regresión Logsítica*, entre otros.
  En particular, se utilizó la función `cv.glmnet()` que ajusta un modelo lineal generalizado, emplea la metodología de *validación cruzada* para encontrar el valor óptimo de λ y grafica el resultado.
  Para más información acerca de la paquetería, así como otras funciones dentro de ella, se le recomienda al lector visitar
  https://cran.r-project.org/web/packages/glmnet/index.html
- `randomForest`
  Esta paquetería se le atribuye a Andy Liaw & Matthew Wiener (2002) y está basada en el código original de Leo Breiman & Adele Cutler (2004) para el lenguaje de programación FORTRAN.
  Se utilizó la función `randomForest()` para implementar el método de *Random Forest*.
  Para más información acerca de la paquetería, así como otras funciones dentro de ella, se le recomienda al lector visitar
  https://cran.r-project.org/web/packages/randomForest/index.html

Anexo C

Código Empleado en R

En este anexo se presenta el código implementado en **R** para los métodos discutidos en esta tesis. Se presentará el código utilizado para cada conjunto de datos abordado en este trabajo.

C.1. Estudio de Cáncer de Mama

C.1.1. Regresión Logística

El siguiente código corresponde al modelo aditivo:

```
err0tr<-c();err1tr<-c();errglobaltr<-c()
err0tst<-c();err1tst<-c();errglobaltst<-c()
betas<-c()
B=1000
set.seed(2022)
for(i in 1:B){
  trainC=createDataPartition(Datos$Y, p=.9, list=FALSE)
  M1Tr=glm(Y~., family=binomial(link=logit), singular.ok=TRUE,
           data=Datos[trainC,])
  betas<-c(betas,length(M1Tr$coefficients))
  pdatap <- predict(M1Tr, newdata = Datos[trainC,], type = "response")
  pdata <- predict(M1Tr, newdata = Datos[-trainC,], type = "response")
  x<-confusionMatrix(data = as.factor(as.numeric(pdatap>0.5)),
        reference = Datos[trainC,]$Y)
  k<-confusionMatrix(data = as.factor(as.numeric(pdata>0.5)),
        reference = Datos[-trainC,]$Y)
  err0tr<-c(err0tr,x$table[2,1]/sum(x$table[,1]))
  err1tr<-c(err1tr,x$table[1,2]/sum(x$table[,2]))
  errglobaltr<-c(errglobaltr,1-x$overall[1])
  err0tst<-c(err0tst,k$table[2,1]/sum(k$table[,1]))
  err1tst<-c(err1tst,k$table[1,2]/sum(k$table[,2]))
  errglobaltst<-c(errglobaltst,1-k$overall[1])
}
```

El siguiente código corresponde al modelo con interacciones dos a dos:

```
err0tr<-c();err1tr<-c();errglobaltr<-c()
err0tst<-c();err1tst<-c();errglobaltst<-c()
betas<-c()
B=1000
set.seed(2020)
for(i in 1:B){
    trainC=createDataPartition(Datos$Y, p=.9, list=FALSE)
    M1Tr=glm(Y~.^2, family=binomial(link=logit), singular.ok=TRUE,
             data=Datos[trainC,])
    betas<-c(betas,length(M1Tr$coefficients))
    pdatap <- predict(M1Tr, newdata = Datos[trainC,], type = "response")
    pdata <- predict(M1Tr, newdata = Datos[-trainC,], type = "response")
    x<-confusionMatrix(data = as.factor(as.numeric(pdatap>0.5)),
            reference = Datos[trainC,]$Y)
    k<-confusionMatrix(data = as.factor(as.numeric(pdata>0.5)),
            reference = Datos[-trainC,]$Y)
    err0tr<-c(err0tr,x$table[2,1]/sum(x$table[,1]))
    err1tr<-c(err1tr,x$table[1,2]/sum(x$table[,2]))
    errglobaltr<-c(errglobaltr,1-x$overall[1])
    err0tst<-c(err0tst,k$table[2,1]/sum(k$table[,1]))
    err1tst<-c(err1tst,k$table[1,2]/sum(k$table[,2]))
    errglobaltst<-c(errglobaltst,1-k$overall[1])
  }
```

C.1.2. *Stepwise Selection*

El siguiente código corresponde al método de *stepwsise selection* utilizando el criterio AIC:

```
err0tr<-c();err1tr<-c();errglobaltr<-c()
err0tst<-c();err1tst<-c();errglobaltst<-c()
betas<-c()
B=1000
set.seed(1999)
for(i in 1:B){
    trainC=createDataPartition(Datos$Y, p=.9, list=FALSE)
    M1Tr=step(glm(Y~.^2, family=binomial(link=logit), singular.ok=TRUE,
             data=Datos[trainC,]),trace=0,k=2)
    betas<-c(betas,length(M1Tr$coefficients))
    pdatap <- predict(M1Tr, newdata = Datos[trainC,], type = "response")
    pdata <- predict(M1Tr, newdata = Datos[-trainC,], type = "response")
    x<-confusionMatrix(data = as.factor(as.numeric(pdatap>0.5)),
            reference = Datos[trainC,]$Y)
    k<-confusionMatrix(data = as.factor(as.numeric(pdata>0.5)),
            reference = Datos[-trainC,]$Y)
    err0tr<-c(err0tr,x$table[2,1]/sum(x$table[,1]))
    err1tr<-c(err1tr,x$table[1,2]/sum(x$table[,2]))
```

```
    errglobaltr<-c(errglobaltr,1-x$overall[1])
    err0tst<-c(err0tst,k$table[2,1]/sum(k$table[,1]))
    err1tst<-c(err1tst,k$table[1,2]/sum(k$table[,2]))
    errglobaltst<-c(errglobaltst,1-k$overall[1])
  }
```

Utilizando el criterio BIC:

```
err0tr<-c();err1tr<-c();errglobaltr<-c()
err0tst<-c();err1tst<-c();errglobaltst<-c()
betas<-c()
B=1000
set.seed(123)
for(i in 1:B){
    trainC=createDataPartition(Datos$Y, p=.9, list=FALSE)
    M1Tr=step(glm(Y~.^2, family=binomial(link=logit), singular.ok=TRUE,
                  data=Datos[trainC,]),trace=0,k=log(nrow(Datos[trainC,])))
    betas<-c(betas,length(M1Tr$coefficients))
    pdatap <- predict(M1Tr, newdata = Datos[trainC,], type = "response")
    pdata <- predict(M1Tr, newdata = Datos[-trainC,], type = "response")
    x<-confusionMatrix(data = as.factor(as.numeric(pdatap>0.5)),
            reference = Datos[trainC,]$Y)
    k<-confusionMatrix(data = as.factor(as.numeric(pdata>0.5)),
            reference = Datos[-trainC,]$Y)
    err0tr<-c(err0tr,x$table[2,1]/sum(x$table[,1]))
    err1tr<-c(err1tr,x$table[1,2]/sum(x$table[,2]))
    errglobaltr<-c(errglobaltr,1-x$overall[1])
    err0tst<-c(err0tst,k$table[2,1]/sum(k$table[,1]))
    err1tst<-c(err1tst,k$table[1,2]/sum(k$table[,2]))
    errglobaltst<-c(errglobaltst,1-k$overall[1])
  }
```

C.1.3. Métodos de Regularización

El siguiente código corresponde al método de *ridge regression*:

```
x=model.matrix(Y ~ .^2 , Datos)
y=Datos$Y
err0tr<-c();err1tr<-c();errglobaltr<-c()
err0tst<-c();err1tst<-c();errglobaltst<-c()
betas<-c()
B=1000
set.seed(1530)
for (i in 1:B) {
    trainC=createDataPartition(Datos$Y, p=.9, list=FALSE)
    MCaux<-cv.glmnet(x[trainC,], y[trainC],family="binomial",
            type.measure="class",nfolds=10,alpha=0)
    betas<-c(betas,MCaux$nzero[which(MCaux$lambda==MCaux$lambda.min)]+1)
    k=t(confusion.glmnet(MCaux,s=MCaux$lambda.min ,newx = x[-trainC,],
```

```
            newy = y[-trainC]))
    t=t(confusion.glmnet(MCaux,s=MCaux$lambda.min,newx=x[trainC,],
            newy=y[trainC]))
    err0tst<-c(err0tst,k[1,2]/sum(k[1,]))
    err1tst<-c(err1tst,k[2,1]/sum(k[2,]))
    errglobaltst<-c(errglobaltst,1-sum(diag(k))/sum(k))
    err0tr<-c(err0tr,t[1,2]/sum(t[1,]))
    err1tr<-c(err1tr,t[2,1]/sum(t[2,]))
    errglobaltr<-c(errglobaltr,1-sum(diag(t))/sum(t))
  }
```

Método de *lasso*:

```
x=model.matrix(Y ~ .^2 , Datos)
y=Datos$Y
err0tr<-c();err1tr<-c();errglobaltr<-c()
err0tst<-c();err1tst<-c();errglobaltst<-c()
betas<-c()
B=1000
set.seed(3001)
for (i in 1:B) {
  trainC=createDataPartition(Datos$Y, p=.9, list=FALSE)
  MCaux<-cv.glmnet(x[trainC,], y[trainC],family="binomial",
        type.measure="class",nfolds=10,alpha=1)
  betas<-c(betas,MCaux$nzero[which(MCaux$lambda==MCaux$lambda.min)]+1)
  k=t(confusion.glmnet(MCaux,s=MCaux$lambda.min ,newx = x[-trainC,],
        newy = y[-trainC]))
  t=t(confusion.glmnet(MCaux,s=MCaux$lambda.min,newx=x[trainC,],
        newy=y[trainC]))
  err0tst<-c(err0tst,k[1,2]/sum(k[1,]))
  err1tst<-c(err1tst,k[2,1]/sum(k[2,]))
  errglobaltst<-c(errglobaltst,1-sum(diag(k))/sum(k))
  err0tr<-c(err0tr,t[1,2]/sum(t[1,]))
  err1tr<-c(err1tr,t[2,1]/sum(t[2,]))
  errglobaltr<-c(errglobaltr,1-sum(diag(t))/sum(t))
}
```

Para el método de *elastic net*, se buscó el valor óptimo de α utilizando el siguiente código:

```
XT=model.matrix(Y~.^2 , Datos)
YT=Datos$Y
errtestfinal<-c()
err0tst<-c()
err1tst<-c()
alpha_1<-c()
set.seed(123)
system.time(
  for (i in 1:1000) {
    trainC=createDataPartition(Datos$Y, p=.9, list=FALSE)
    x=model.matrix(Y~.^2 , Datos[trainC,])
```

```
    y=Datos[trainC,]$Y
    errorvalidation<-c()
    for (alpha in seq(0.1,0.9,by=0.1)) {
      train=createDataPartition(Datos[trainC,]$Y, p=.8, list=FALSE)
      MCaux<-cv.glmnet(x[train,], y[train],family="binomial",
      type.measure="class",nfolds=10,alpha=alpha)
      k=t(confusion.glmnet(MCaux,s=MCaux$lambda.min ,newx = x[-train,],
      newy = y[-train]))
      errorvalidation<-c(errorvalidation,1-sum(diag(k))/sum(k))
    }
    alpha_1<-c(alpha_1,which.min(errorvalidation)/10)
    Elastic<-cv.glmnet(XT[trainC,], YT[trainC],family="binomial",
    type.measure="class",nfolds=10,alpha=which.min(errorvalidation)/10)
    betas<-c(betas,Elastic$nzero[which(Elastic$lambda==Elastic$lambda.min)]+1)
    k=t(confusion.glmnet(Elastic,s=Elastic$lambda.min ,
    newx = XT[-trainC,], newy = YT[-trainC]))
    errtestfinal<-c(errtestfinal,1-sum(diag(k))/sum(k))
    err0tst<-c(err0tst,k[1,2]/sum(k[1,]))
    err1tst<-c(err1tst,k[2,1]/sum(k[2,]))
    if(i%%100==0){print(i)}
  }
)
```

C.1.4. Análisis Discriminante

El siguiente código corresponde al método LDA:

```
err0tr<-c();err1tr<-c();errglobaltr<-c()
err0tst<-c();err1tst<-c();errglobaltst<-c()
B=1000
set.seed(3)
for(i in 1:B){
    trainC=createDataPartition(Datos$Y, p=.9, list=FALSE)
    ldatrain=lda(Y ~ ., data=Datos[trainC,])
    pdatap <- predict(ldatrain,newdata = Datos[trainC,],
            type="class")$class
    pdata <- predict(ldatrain, newdata = Datos[-trainC,],
            type = "class")$class
    x<-confusionMatrix(data = pdatap, reference = Datos[trainC,]$Y)
    k<-confusionMatrix(data = pdata, reference = Datos[-trainC,]$Y)
    err0tr<-c(err0tr,x$table[2,1]/sum(x$table[,1]))
    err1tr<-c(err1tr,x$table[1,2]/sum(x$table[,2]))
    errglobaltr<-c(errglobaltr,1-x$overall[1])
    err0tst<-c(err0tst,k$table[2,1]/sum(k$table[,1]))
    err1tst<-c(err1tst,k$table[1,2]/sum(k$table[,2]))
    errglobaltst<-c(errglobaltst,1-k$overall[1])
  }
```

Métodología QDA:

```
err0tr<-c();err1tr<-c();errglobaltr<-c()
err0tst<-c();err1tst<-c();errglobaltst<-c()
B=1000
set.seed(22)
for(i in 1:B){
    trainC=createDataPartition(Datos$Y, p=.9, list=FALSE)
    q<-qda(Y ~ ., data=Datos[trainC,])
    pdatap <- predict(q,newdata = Datos[trainC,],
            type="class")$class
    pdata <- predict(q, newdata = Datos[-trainC,],
            type = "class")$class
    x<-confusionMatrix(data = pdatap, reference = Datos[trainC,]$Y)
    k<-confusionMatrix(data = pdata, reference = Datos[-trainC,]$Y)
    err0tr<-c(err0tr,x$table[2,1]/sum(x$table[,1]))
    err1tr<-c(err1tr,x$table[1,2]/sum(x$table[,2]))
    errglobaltr<-c(errglobaltr,1-x$overall[1])
    err0tst<-c(err0tst,k$table[2,1]/sum(k$table[,1]))
    err1tst<-c(err1tst,k$table[1,2]/sum(k$table[,2]))
    errglobaltst<-c(errglobaltst,1-k$overall[1])
  }
```

C.1.5. *Random Forest*

Para el método de *random forest*, primero se buscó la mejor configuración para el parámetro *mtry*. Se utilizó el siguiente código:

```
errtestfinal<-c()
err0tst<-c()
err1tst<-c()
mtry_1<-c()
set.seed(123)
system.time(
  for (i in 1:1000) {
    trainC=createDataPartition(Datos$Y, p=.9, list=FALSE)
    temp<-Datos[trainC,]
    errorvalidation<-c()
    for (mtry in 1:9) {
      train=createDataPartition(temp$Y, p=.8, list=FALSE)
      fit2=randomForest(Y~., data=temp[train,], mtry=mtry, ntree=1000)
      pdata <- predict(fit2, newdata = temp[-train,], type = "class")
      x<-confusionMatrix(data = pdata, reference = temp[-train,]$Y)
      errorvalidation<-c(errorvalidation,1-x$overall[1])
    }
    mtry_1<-c(mtry_1,which.min(errorvalidation))
    RF<-randomForest(Y~., data=Datos[trainC,],
    mtry=which.min(errorvalidation), ntree=1000)
    k<-confusionMatrix(data = predict(RF, newdata = Datos[-trainC,],
    type = "class"), reference = Datos[-trainC,]$Y)
```

```
    errtestfinal<-c(errtestfinal,1-x$overall[1])
    err0tst<-c(err0tst,k$table[2,1]/sum(k$table[,1]))
    err1tst<-c(err1tst,k$table[1,2]/sum(k$table[,2]))
    if(i%%100==0){print(i)}
  }
)
```

C.2. Datos Simulados

Para los datos simulados se utilizó el siguiente código de Eslava, G. & Pérez, G. (2022):

```
rho=c(.3,-.3)
p.1=c(.5,.5)
medias=c(0,0)
nCat=4 #número de variables categóricas
nCont=6 #número de variables continuas
Probs.cbin=function(rho, p1){
  p11=rho*(p1-(p1)^2)+ (p1)^2
  p01=p1-p11
  p10=p01
  p00=1-p11-2*p01
  return(c(p00,p01,p10,p11))
}
###Funciones auxiliares
#Probabilidades para los cliques de la grafica en el caso binario
#Se define una correlacion rho que sera la misma que se usara para
#las continuas.
#Sea pkl la probabilidad de la variable X_i de tomar el valor k
#    y la X_j el valor l, donde k, l \in {0,1}.
#Basta con definir la probabilidad marginal p1=p(X_j=1) para tener
#p11=rho (p_1-p_1^2)+ p_1^2 y así definir las probabilidades del clique

Probs.cbin=function(rho, p1){
  p11=rho*(p1-(p1)^2)+ (p1)^2
  p01=p1-p11
  p10=p01
  p00=1-p11-2*p01
  return(c(p00,p01,p10,p11))
}

sigmaAR1=function(rho, p)
{
  sigmas=matrix(0,p,p)
  for(i in 1:p){
    for(j in 1:p){
      sigmas[i,j]=rho^(abs(j-i))
      sigmas[j,i]=rho^(abs(j-i))
    }
```

```
  }
  return(sigmas)
}

pro0=Probs.cbin(rho[1], p.1[1])
pro1=Probs.cbin(rho[2], p.1[2])
mu0=medias[1]*rep(1,nCont)
mu1=medias[2]*rep(1,nCont)
sigma0=sigmaAR1(rho[1], nCont)
sigma1=sigmaAR1(rho[2], nCont)

#Funcion para simular datos
SimPathPath=function(nsim, rho, p.1, medias, nCat, nCont){
  prob=Probs.cbin(rho, p.1)
  mu=medias*rep(1,nCont)
  sigma=sigmaAR1(rho, nCont)
  Datos=as.data.frame(matrix(0, nrow=nsim, ncol=nCat+nCont))
  Datos[,1]=rbinom(nsim,1,p.1)
  for(jk in 2:nCat){
    Datos[,jk]=rbinom(nsim,1, prob[2]/(1-p.1)*(Datos[,jk-1]==0)
                +prob[4]/(p.1)*(Datos[,jk-1]==1))
  }
  Datos[,(nCat+1):(nCat+nCont)]=rmvnorm(nsim,mu,sigma)
  return(Datos)
}

nsim0=100
nsim1=100
SimDatos=function(nsim0,nsim1){
  Datos0=SimPathPath(nsim0, rho[1], p.1[1], medias[1], nCat, nCont)
  Datos0$Y=0
  Datos1=SimPathPath(nsim1, rho[2], p.1[2], medias[2], nCat, nCont)
  Datos1$Y=1
  Train=rbind(Datos0, Datos1)
  for(i in c(1:4,11)){
    Train[,i]=as.factor(Train[,i])
  }
  return(Train)
}
```

El siguiente código corresponde al modelo aditivo:

```
err0tr<-c();err1tr<-c();errglobaltr<-c()
err0tst<-c();err1tst<-c();errglobaltst<-c()
betas<-c()
B=1000
set.seed(2022)
for(i in 1:B){
    Train=SimDatos(nsim0,nsim1)
```

```
    Test=SimDatos(nsim0,nsim1)
    M1Tr=glm(Y~., family=binomial(link=logit), singular.ok=TRUE,
             data=Train)
    betas<-c(betas,length(M1Tr$coefficients))
    pdatap <- predict(M1Tr, newdata = Train, type = "response")
    pdata <- predict(M1Tr, newdata = Test, type = "response")
    x<-confusionMatrix(data = as.factor(as.numeric(pdatap>0.5)),
            reference = Train$Y)
    k<-confusionMatrix(data = as.factor(as.numeric(pdata>0.5)),
            reference = Test$Y)
    err0tr<-c(err0tr,x$table[2,1]/sum(x$table[,1]))
    err1tr<-c(err1tr,x$table[1,2]/sum(x$table[,2]))
    errglobaltr<-c(errglobaltr,1-x$overall[1])
    err0tst<-c(err0tst,k$table[2,1]/sum(k$table[,1]))
    err1tst<-c(err1tst,k$table[1,2]/sum(k$table[,2]))
    errglobaltst<-c(errglobaltst,1-k$overall[1])
  }
```

Para el modelo con interacciones dos a dos:

```
err0tr<-c();err1tr<-c();errglobaltr<-c()
err0tst<-c();err1tst<-c();errglobaltst<-c()
betas<-c()
B=1000
set.seed(2020)
for(i in 1:B){
    Train=SimDatos(nsim0,nsim1)
    Test=SimDatos(nsim0,nsim1)
    M1Tr=glm(Y~.^2, family=binomial(link=logit), singular.ok=TRUE,
             data=Train)
    betas<-c(betas,length(M1Tr$coefficients))
    pdatap <- predict(M1Tr, newdata = Train, type = "response")
    pdata <- predict(M1Tr, newdata = Test, type = "response")
    x<-confusionMatrix(data = as.factor(as.numeric(pdatap>0.5)),
            reference = Train$Y)
    k<-confusionMatrix(data = as.factor(as.numeric(pdata>0.5)),
            reference = Test$Y)
    err0tr<-c(err0tr,x$table[2,1]/sum(x$table[,1]))
    err1tr<-c(err1tr,x$table[1,2]/sum(x$table[,2]))
    errglobaltr<-c(errglobaltr,1-x$overall[1])
    err0tst<-c(err0tst,k$table[2,1]/sum(k$table[,1]))
    err1tst<-c(err1tst,k$table[1,2]/sum(k$table[,2]))
    errglobaltst<-c(errglobaltst,1-k$overall[1])
  }
```

C.2.1. *Stepwise Selection*

El siguiente código corresponde al criterio AIC:

```
err0tr<-c();err1tr<-c();errglobaltr<-c()
err0tst<-c();err1tst<-c();errglobaltst<-c()
betas<-c()
B=1000
set.seed(1999)
for(i in 1:B){
    Train=SimDatos(nsim0,nsim1)
    Test=SimDatos(nsim0,nsim1)
    M1Tr=step(glm(Y~.^2, family=binomial(link=logit), singular.ok=TRUE,
                  data=Train),trace=0,k=2)
    betas<-c(betas,length(M1Tr$coefficients))
    pdatap <- predict(M1Tr, newdata = Train, type = "response")
    pdata <- predict(M1Tr, newdata = Test, type = "response")
    x<-confusionMatrix(data = as.factor(as.numeric(pdatap>0.5)),
            reference = Train$Y)
    k<-confusionMatrix(data = as.factor(as.numeric(pdata>0.5)),
            reference = Test$Y)
    err0tr<-c(err0tr,x$table[2,1]/sum(x$table[,1]))
    err1tr<-c(err1tr,x$table[1,2]/sum(x$table[,2]))
    errglobaltr<-c(errglobaltr,1-x$overall[1])
    err0tst<-c(err0tst,k$table[2,1]/sum(k$table[,1]))
    err1tst<-c(err1tst,k$table[1,2]/sum(k$table[,2]))
    errglobaltst<-c(errglobaltst,1-k$overall[1])
  }
```

Para el criterio BIC se utilizó el siguiente código:

```
err0tr<-c();err1tr<-c();errglobaltr<-c()
err0tst<-c();err1tst<-c();errglobaltst<-c()
betas<-c()
B=1000
set.seed(123)
for(i in 1:B){
    Train=SimDatos(nsim0,nsim1)
    Test=SimDatos(nsim0,nsim1)
    M1Tr=step(glm(Y~.^2, family=binomial(link=logit), singular.ok=TRUE,
                  data=Train),trace=0,k=log(nrow(Train)))
    betas<-c(betas,length(M1Tr$coefficients))
    pdatap <- predict(M1Tr, newdata = Train, type = "response")
    pdata <- predict(M1Tr, newdata = Test, type = "response")
    x<-confusionMatrix(data = as.factor(as.numeric(pdatap>0.5)),
            reference = Train$Y)
    k<-confusionMatrix(data = as.factor(as.numeric(pdata>0.5)),
            reference = Test$Y)
    err0tr<-c(err0tr,x$table[2,1]/sum(x$table[,1]))
    err1tr<-c(err1tr,x$table[1,2]/sum(x$table[,2]))
    errglobaltr<-c(errglobaltr,1-x$overall[1])
    err0tst<-c(err0tst,k$table[2,1]/sum(k$table[,1]))
    err1tst<-c(err1tst,k$table[1,2]/sum(k$table[,2]))
```

```
    errglobaltst<-c(errglobaltst,1-k$overall[1])
  }
```

C.2.2. Métodos de Regularización

El método de *ridge regression* fue empleado en el siguiente código:

```
err0tr<-c();err1tr<-c();errglobaltr<-c()
err0tst<-c();err1tst<-c();errglobaltst<-c()
betas<-c()
B=1000
set.seed(200)
  for (i in 1:B) {
    Train=SimDatos(nsim0,nsim1)
    Test=SimDatos(nsim0,nsim1)
    xtrain=model.matrix(Y ~ .^2 , Train)
    ytrain=Train$Y
    xtest=model.matrix(Y ~ .^2 , Test)
    ytest=Test$Y
    MCaux<-cv.glmnet(xtrain, ytrain,family="binomial",
            type.measure="class",nfolds=10,alpha=0)
    betas<-c(betas,MCaux$nzero[which(MCaux$lambda==MCaux$lambda.min)]+1)
    k=t(confusion.glmnet(MCaux,s=MCaux$lambda.min ,newx = xtest,
        newy = ytest))
    t=t(confusion.glmnet(MCaux,s=MCaux$lambda.min,newx=xtrain,
        newy=ytrain))
    err0tst<-c(err0tst,k[1,2]/sum(k[1,]))
    err1tst<-c(err1tst,k[2,1]/sum(k[2,]))
    errglobaltst<-c(errglobaltst,1-sum(diag(k))/sum(k))
    err0tr<-c(err0tr,t[1,2]/sum(t[1,]))
    err1tr<-c(err1tr,t[2,1]/sum(t[2,]))
    errglobaltr<-c(errglobaltr,1-sum(diag(t))/sum(t))
  }
```

Para *lasso* se implementó el siguiente código:

```
err0tr<-c();err1tr<-c();errglobaltr<-c()
err0tst<-c();err1tst<-c();errglobaltst<-c()
betas<-c()
B=1000
set.seed(3001)
  for (i in 1:B) {
    Train=SimDatos(nsim0,nsim1)
    Test=SimDatos(nsim0,nsim1)
    xtrain=model.matrix(Y ~ .^2 , Train)
    ytrain=Train$Y
    xtest=model.matrix(Y ~ .^2 , Test)
    ytest=Test$Y
    MCaux<-cv.glmnet(xtrain, ytrain,family="binomial",
```

```
            type.measure="class",nfolds=10,alpha=1)
    betas<-c(betas,MCaux$nzero[which(MCaux$lambda==MCaux$lambda.min)]+1)
    k=t(confusion.glmnet(MCaux,s=MCaux$lambda.min ,newx = xtest,
        newy = ytest))
    t=t(confusion.glmnet(MCaux,s=MCaux$lambda.min,newx=xtrain,
        newy=ytrain))
    err0tst<-c(err0tst,k[1,2]/sum(k[1,]))
    err1tst<-c(err1tst,k[2,1]/sum(k[2,]))
    errglobaltst<-c(errglobaltst,1-sum(diag(k))/sum(k))
    err0tr<-c(err0tr,t[1,2]/sum(t[1,]))
    err1tr<-c(err1tr,t[2,1]/sum(t[2,]))
    errglobaltr<-c(errglobaltr,1-sum(diag(t))/sum(t))
  }
```

Para *elastic net* se buscó el valor óptimo de α utilizando el siguiente código:

```
errelasticex<-c()
set.seed(123456)
for (alpha in seq(0.1,0.9,0.1)) {
    errglobaltst<-c()
    for (i in 1:1000) {
        Train=SimDatos(nsim0,nsim1)
        Test=SimDatos(nsim0,nsim1)
        x=model.matrix(Y~.^2 , Train)
        y=Train$Y
        xtest=model.matrix(Y ~ .^2 , Test)
        ytest=Test$Y
        MCaux<-cv.glmnet(x, y,family="binomial",type.measure="class",
        nfolds=10,alpha=alpha)
        k=t(confusion.glmnet(MCaux,s=MCaux$lambda.min ,newx = xtest,
            newy = ytest))
        errglobaltst<-c(errglobaltst,1-sum(diag(k))/sum(k))
}
errelasticex<-c(errelasticex,mean(errglobaltst))
}
```

Finalmente, se empleó el siguiente código:

```
err0tr<-c();err1tr<-c();errglobaltr<-c()
err0tst<-c();err1tst<-c();errglobaltst<-c()
betas<-c()
B=1000
set.seed(200)
for (i in 1:B) {
    Train=SimDatos(nsim0,nsim1)
    Test=SimDatos(nsim0,nsim1)
    xtrain=model.matrix(Y ~ .^2 , Train)
    ytrain=Train$Y
    xtest=model.matrix(Y ~ .^2 , Test)
    ytest=Test$Y
```

```
    MCaux<-cv.glmnet(xtrain, ytrain,family="binomial",
    type.measure="class",nfolds=10,alpha=0.6)
    betas<-c(betas,MCaux$nzero[which(MCaux$lambda==MCaux$lambda.min)]+1)
    k=t(confusion.glmnet(MCaux,s=MCaux$lambda.min ,newx = xtest,
        newy = ytest))
    t=t(confusion.glmnet(MCaux,s=MCaux$lambda.min,newx=xtrain,
        newy=ytrain))
    err0tst<-c(err0tst,k[1,2]/sum(k[1,]))
    err1tst<-c(err1tst,k[2,1]/sum(k[2,]))
    errglobaltst<-c(errglobaltst,1-sum(diag(k))/sum(k))
    err0tr<-c(err0tr,t[1,2]/sum(t[1,]))
    err1tr<-c(err1tr,t[2,1]/sum(t[2,]))
    errglobaltr<-c(errglobaltr,1-sum(diag(t))/sum(t))
}
```

C.2.3. Análisis Discriminante

Para LDA se utilizó el siguiente código:

```
err0tr<-c();err1tr<-c();errglobaltr<-c()
err0tst<-c();err1tst<-c();errglobaltst<-c()
B=1000
set.seed(3)
for(i in 1:B){
    Train=SimDatos(nsim0,nsim1)
    Test=SimDatos(nsim0,nsim1)
    ldatrain=lda(Y ~ ., data=Train)
    pdatap <- predict(ldatrain,newdata = Train,type="class")$class
    pdata <- predict(ldatrain, newdata = Test, type = "class")$class
    x<-confusionMatrix(data = pdatap, reference = Train$Y)
    k<-confusionMatrix(data = pdata, reference = Test$Y)
    err0tr<-c(err0tr,x$table[2,1]/sum(x$table[,1]))
    err1tr<-c(err1tr,x$table[1,2]/sum(x$table[,2]))
    errglobaltr<-c(errglobaltr,1-x$overall[1])
    err0tst<-c(err0tst,k$table[2,1]/sum(k$table[,1]))
    err1tst<-c(err1tst,k$table[1,2]/sum(k$table[,2]))
    errglobaltst<-c(errglobaltst,1-k$overall[1])
  }
```

Para el caso de QDA se implementó el siguiente código:

```
err0tr<-c();err1tr<-c();errglobaltr<-c()
err0tst<-c();err1tst<-c();errglobaltst<-c()
B=1000
set.seed(22)
for(i in 1:B){
    Train=SimDatos(nsim0,nsim1)
    Test=SimDatos(nsim0,nsim1)
    q<-qda(Y ~ ., data=Train)
```

```
    pdatap <- predict(q,newdata = Train,type="class")$class
    pdata <- predict(q, newdata = Test, type = "class")$class
    x<-confusionMatrix(data = pdatap, reference = Train$Y)
    k<-confusionMatrix(data = pdata, reference = Test$Y)
    err0tr<-c(err0tr,x$table[2,1]/sum(x$table[,1]))
    err1tr<-c(err1tr,x$table[1,2]/sum(x$table[,2]))
    errglobaltr<-c(errglobaltr,1-x$overall[1])
    err0tst<-c(err0tst,k$table[2,1]/sum(k$table[,1]))
    err1tst<-c(err1tst,k$table[1,2]/sum(k$table[,2]))
    errglobaltst<-c(errglobaltst,1-k$overall[1])
  }
```

C.2.4. *Random Forest*

Para *random forest*, primero se buscó los valores óptimos para *mtry* y *ntree* con el siguiente código:

```
erroresRFExp<- matrix(NA,nrow = 10,ncol = 10)
for(mtry in 1:10){
  set.seed(123+mtry)
  for (ntree in c(100,300,500,1000)) {
    prom<-c()
    for (i in 1:1000){
      Train=SimDatos(nsim0,nsim1)
      Test=SimDatos(nsim0,nsim1)
      fit2=randomForest(Y~., data=Train, mtry=mtry, ntree=ntree)
      pdata <- predict(fit2, newdata = Test, type = "class")
      x<-confusionMatrix(data = pdata, reference = Test$Y)
      prom[i]=1-x$overall[1]}
    erroresRFExp[mtry,ntree/100]=mean(prom)
  }
}
erroresRFExp=erroresRFExp[,-c(2,4,6,7,8,9)]
colnames(erroresRFExp)=c("100","300","500","1000")
```

Finalmente, se utilizó el siguiente código:

```
err0tr<-c()
err1tr<-c()
errglobaltr<-c()
err0tst<-c()
err1tst<-c()
errglobaltst<-c()
B=1000
set.seed(2356)
system.time(
  for(i in 1:B){
    Train=SimDatos(nsim0,nsim1)
    Test=SimDatos(nsim0,nsim1)
```

```
  q=randomForest(Y~., data=Train, mtry=2, ntree=1000)
  pdatap <- predict(q, newdata = Train, type = "class")
  pdata <- predict(q, newdata = Test, type = "class")
  x<-confusionMatrix(data = pdatap, reference = Train$Y)
  k<-confusionMatrix(data = pdata, reference = Test$Y)
  err0tr<-c(err0tr,x$table[2,1]/sum(x$table[,1]))
  err1tr<-c(err1tr,x$table[1,2]/sum(x$table[,2]))
  errglobaltr<-c(errglobaltr,1-x$overall[1])
  err0tst<-c(err0tst,k$table[2,1]/sum(k$table[,1]))
  err1tst<-c(err1tst,k$table[1,2]/sum(k$table[,2]))
  errglobaltst<-c(errglobaltst,1-k$overall[1])
  if(i%%100==0) print(i)
})
```

Bibliografía

[1] Anderson, T. W. (2009). *An Introduction to Multivariate Statistical Analysis*. (3rd ed.). Wiley.

[2] Armitage, P., McPherson, C.K. & Copas, J.C.(1969). Statistical Studies of Prognosis in Advanced Breast Cancer. *Journal of Chronic Diseases, 22*(5), 343-360. https://doi.org/10.1016/0021-9681(69)90076-9

[3] Bartlett, M.S. & Please, N.W. (1963). Discrimination in the case of zero mean differences. *Biometrika, 50*(1-2), 17–21. doi:10.1093/biomet/50.1-2.17

[4] Bishop, C.M. (2006). *Pattern Recognition and Machine Learning*. Springer.

[5] Cortés, L.C. (2022). *Boosting con árboles de decisión y random forest*. Tesis de licenciatura. Universidad Nacional Autónoma de México.

[6] Efron, B. & Hastie, T. (2021). *Computer Age Statistical Inference, Student Edition: Algorithms, Evidence, and Data Science*. Cambridge University Press.

[7] Eslava, G. & Pérez, G. (2022). Classification using binary and continuous variables. In *Symposium i anvendt statistik*, SEGES, Landbrug & Fødevarer, 2022, pp. 14-22

[8] Fisher, R.A. (1936). The use of Multiple Measurements in Taxonomic problems. *Annals of Eugenics, 7*(2), 179-188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x

[9] Hastie, T., Tibshirani, R. & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. (2nd ed.). Springer.

[10] Hastie, T., Tibshirani, R., & Wainwright, M. (2020). *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC Press.

[11] Hosmer, D.W., Lemeshow, S., & Sturdivant, R.X. (2013). *Applied Logistic Regression* (3rd ed.). Wiley.

[12] James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R (Springer Texts in Statistics)* (2nd ed.). Springer Publishing.

[13] Krzanowski, W.J. (1976). Canonical Representation of the Location Model for Discrimination or Classification. *Journal of the American Statistical Association, 71*(356), 845. https://doi.org/10.2307/2286849

[14] Lauritzen, S.L. (1996). *Graphical Models*. Clarendon Press, Oxford.

[15] R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[16] Ripley, B.D. (1994). Neural Networks and Related Methods for Classification. *Journal of the Royal Statistical Society. Series B (Methodological), 56*(3), 409–456. http://www.jstor.org/stable/2346118

[17] Venables, W.N., & Ripley, B.D. (2002). *Modern Applied Statistics with S* (4th ed.). Springer, New York

[18] Zou, H. & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*(2), 301-320. https://doi.org/10.1111/j.1467-9868.2005.00503.x