UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
POSGRADO EN CIENCIA E INGENIERÍA DE LA COMPUTACIÓN
INSTITUTO DE INVESTIGACIONES EN MATEMÁTICAS APLICADAS Y SISTEMAS

ESTUDIO DE PATRONES DE EVOLUCIÓN DE HÁBITOS EN TRABAJADORES DE LA UNAM, CON APLICACIÓN EN LA PREDICCIÓN DE OBESIDAD

TESIS

QUE PARA OPTAR POR EL GRADO DE:
MAESTRO EN CIENCIAS DE LA COMPUTACIÓN

PRESENTA:
VICENTE ALBÍTER ALPÍZAR

DIRECTOR DE TESIS:
DR. CHRISTOPHER R. STEPHENS
Posgrado en Ciencia e Ingeniería de la Computación

CIUDAD UNIVERSITARIA, CDMX
MAYO 2021

Contenidos

Capítulo 1. Introducción 3

Capítulo 2. Planteamiento del problema 6

2.1 - Descripción de la base de datos 6

2.2 – Objetivos del estudio 11

Capítulo 3. Evolución de hábitos en el tiempo 12

3.1 – Preprocesamiento de las variables de interés 12

3.2 – Análisis de la evolución de las variables de *Autoevaluación de Salud* en el tiempo 20

Capítulo 4. Relación entre historias y subgrupos de la población 48

4.1 – Análisis de épsilons asociados a historiales de hábitos 48

Capítulo 5. Modelos de Predicción 67

5.1 – Herramientas para la construcción de los modelos de predicción 67

5.2 – Metodología del entrenamiento y evaluación de los modelos de predicción 76

5.3 – Predicción de Obesidad a partir de historiales de Ejercicio 77

5.4 – (Anexo) Otros Resultados 84

Capítulo 6. Conclusiones 90

Bibliografía 94

Capítulo 1. Introducción

La obesidad es una condición compleja, asociada a múltiples factores de riesgo, tales como tener hábitos de vida poco saludable (vida sedentaria, dieta descontrolada, poco sueño, etcétera), la edad, la raza o etnicidad, el sexo, la condición socioeconómica y la genética [1][2]. Las complicaciones de salud que pueden generarse como consecuencia de esta afección son, entre otras, síndrome metabólico, presión arterial alta, ateroesclerosis, enfermedad del corazón, diabetes, niveles elevados de colesterol en sangre, distintos tipos de cáncer y trastornos del sueño.

En México, alrededor del 72.5% de los adultos del país tiene sobrepeso. Sólo en 2014, los costos médicos derivados de la obesidad fueron estimados en 151 894 millones de pesos, lo cual equivale a 34% del gasto público en salud. En términos de productividad, la pérdida económica del país debido a esta condición se estima en 71 669 millones de pesos (0.4% del PIB) por año [3]. En un análisis de 2004 sobre la carga de la obesidad en el país, se encontró que alrededor del 25.3% del total de las muertes en el país se derivaron de esta afección [4].

Con este contexto, resulta de gran relevancia para la ciencia del país el ayudar a desarrollar estrategias de prevención de obesidad basadas en el entendimiento de sus causas. Con esta motivación, en el artículo de 2018 titulado *Bayesian Classification of Personal Histories – An application to the Obesity Epidemic*, Stephens et al. vincularon el historial de ejercicio (últimos 20 años) de trabajadores de la Universidad Nacional Autónoma de México (UNAM) con su obesidad actual [5]. Entre otras cosas se encontró que, contrario a la intuición, la obesidad parece estar más relacionada con la pérdida de un buen hábito de ejercicio que con el mantenimiento de un mal hábito de ejercicio a lo largo del tiempo muestreado. El presente trabajo de tesis está motivado por este estudio y está basado en la misma base de datos sobre la que dicho estudio se construyó.

En términos prácticos, este trabajo es una extensión a más grande escala del estudio referenciado en [5]. Tomamos como meta última la predicción de la obesidad a partir del conocimiento de los historiales de hábitos de las personas de la base de datos, pero antes de ello hacemos una extensa exploración al respecto de la evolución de dichos historiales en sí, y de la relación de ciertos historiales con ciertos grupos de personas. Para hacer esto presentamos el contenido en la siguiente secuencia:

- En el capítulo 2 se plantea el problema y se describe a detalle las variables de interés (historiales de hábitos) dentro de la base de datos utilizada. Con estas definiciones, se plantean los objetivos generales del estudio.

- En el capítulo 3 describimos un preprocesamiento de las variables de interés, de manera que éstas puedan ser manipuladas de forma más sencilla. Asimismo, establecemos una serie de mecanismos para describir la forma en que dichas variables evolucionan en el tiempo en diferentes grupos de personas, divididos según su nivel de estudios, su ocupación, su sexo y según si padecen o no de obesidad. La idea tras explorar todas estas divisiones es la de encontrar si el grupo de personas con obesidad muestra patrones distintos de sus variables de interés con respecto a otros grupos de personas en donde la obesidad de la gente no es relevante.

- En el capítulo 4 describimos una serie de nuevos mecanismos para estudiar y cuantificar la relación estadística entre las variables de interés y diferentes grupos de personas (usando las mismas divisiones de personas usadas en el capítulo 3). Esto con el fin de encontrar patrones interesantes (aquellos con una relación estadísticamente significativa) asociados al grupo de personas que padece obesidad, y a aquellos grupos cuya condición de obesidad es irrelevante.

- En el capítulo 5 retomamos algunos mecanismos planteados en los capítulos 3 y 4, e introducimos una serie de modelos bayesianos de clasificación con el fin de construir modelos de predicción de obesidad basados en las variables de interés. La justificación de usar mecanismos bayesianos y no otro tipo de algoritmos de aprendizaje supervisado para esta aplicación es que estos son altamente interpretables (pues conocemos con exactitud la distribución de probabilidad que un clasificador del tipo *Naive Bayes* aprende antes de asignar una clase). Además, entre otras cosas, pueden asignar *scores* numéricos a las entradas y dichos *scores* están monótonamente correlacionados con la probabilidad de que estas entradas pertenezcan o no a una clase [6] (pues por definición, el principio de asignación de clase de un clasificador del tipo *Naive Bayes* se basa en la estimación de las probabilidades posteriores, una vez aprendida la distribución de probabilidad en cuestión). Esta asignación de *scores* permite no sólo la clasificación de individuos, sino también del ordenamiento de estos. La utilidad de esta característica será introducida y explicada más adelante.

La idea de tener a la predicción de obesidad como meta última es la de la *accionabilidad* que eso le confiere a este trabajo: si podemos entender qué tipo de perfiles de personas

están asociados a qué tipos de historiales de hábitos y, además, sabemos qué tipo de historiales de hábitos están asociados a la obesidad, los modelos desarrollados por el presente trabajo podrán ayudar a determinar qué trabajadores de la UNAM son más propensos a padecer obesidad, a partir de sus historiales de hábitos. Tener esta clase de perspectiva puede ayudar a hacer campañas de prevención y de mejoramiento de la salud mejor orientadas (pues se tendrá ahora una mejor idea de en qué grupos de personas habrá de centrarse en particular) y enfocadas (pues habiendo determinado la causalidad entre ciertos historiales de hábitos y ciertos padecimientos, se podrá determinar con más exactitud qué tipo de personas son las que deben recibir más atención).

Capítulo 2. Planteamiento del problema

2.1 - Descripción de la base de datos

La base de datos referenciada en el estudio [1] contiene información respecto a los datos personales, antropometría, antecedentes familiares, antecedentes personales, datos de laboratorio, cuestionario de autoevaluación, nutrición y estilo de vida de 1076 trabajadores académicos y no académicos de la UNAM [5]. Dicha base de datos fue recolectada bajo el proyecto FM/DI/023/2014, usando protocolos aprobados por el Comité de ética de la Facultad de Medicina de la UNAM, y fue proporcionada al autor de este trabajo por el Dr. Christopher Stephens (C3, UNAM). Cada entrada de la base de datos (identificada por un número de folio único) representa pues a cada uno de estos 1076 trabajadores. En particular, el alcance de este trabajo se centra en el estudio de ciertas variables pertenecientes a los conjuntos de variables que en la base de datos están etiquetados como "*Variables de Autoevaluación de Salud*", "*Variables de Antropometría*" y "*Variables de Datos Personales*".

Variables de "Autoevaluación de Salud"

Los datos contenidos dentro de las variables de este subconjunto son todos resultado de un cuestionario de autoevaluación que se le aplicó a todos los participantes de la encuesta sobre la que se construyó esta base de datos. Estas variables son de contenido histórico y no comprobable y, en este sentido, comparten la particularidad de que todas representan datos cuya veracidad depende completamente de la honestidad y el criterio del encuestado. Las variables de interés que caen dentro de este subconjunto se describen a continuación.

1. Ejercicio: Son las variables que representan la cantidad de ejercicio semanal, en horas, reportado por el sujeto en cuestión, para varios tiempos de referencia. Las etiquetas asociadas a este conjunto de variables son las siguientes:
   - *ejer_act*: Cantidad de horas semanales de ejercicio que el sujeto en cuestión realiza actualmente.
   - *ejer1:* Cantidad de horas semanales de ejercicio que el sujeto en cuestión realizaba hace 1 año.
   - *ejer5*: Cantidad de horas semanales de ejercicio que el sujeto en cuestión realizaba hace 5 años.
   - *ejer10:* Cantidad de horas semanales de ejercicio que el sujeto en cuestión realizaba hace 10 años.

- *ejer20:* Cantidad de horas semanales de ejercicio que el sujeto en cuestión realizaba hace 20 años.
- *ejer30:* Cantidad de horas semanales de ejercicio que el sujeto en cuestión realizaba hace 30 años.

2. Condición Física: Son las variables que representan el estado de condición física del sujeto en cuestión, para varios tiempos de referencia. Las etiquetas asociadas a este conjunto de variables son las siguientes:
   - *condi_act*: Estado de condición física en que el sujeto en cuestión se encuentra actualmente, de acuerdo con los valores mostrados en la Tabla 2.1.
   - *condi1:* Estado de condición física en que el sujeto en cuestión se encontraba hace 1 año.
   - *condi5*: Estado de condición física en que el sujeto en cuestión se encontraba hace 5 años.
   - *condi10:* Estado de condición física en que el sujeto en cuestión se encontraba hace 10 años.
   - *condi20:* Estado de condición física en que el sujeto en cuestión se encontraba hace 20 años.
   - *condi30:* Estado de condición física en que el sujeto en cuestión se encontraba hace 30 años.

3. Salud: Son las variables que representan el estado de salud reportado por el sujeto en cuestión, para varios tiempos de referencia. Las etiquetas asociadas a este conjunto de variables son las siguientes:
   - *salud_act*: Estado de salud en que el sujeto en cuestión se encuentra actualmente, de acuerdo con los valores mostrados en la Tabla 2.1.
   - *salud1:* Estado de salud en que el sujeto en cuestión se encontraba hace 1 año.
   - *salud5*: Estado de salud en que el sujeto en cuestión se encontraba hace 5 años.
   - *salud10:* Estado de salud en que el sujeto en cuestión se encontraba hace 10 años.
   - *salud20:* Estado de salud en que el sujeto en cuestión se encontraba hace 20 años.
   - *salud30:* Estado de salud en que el sujeto en cuestión se encontraba hace 30 años.

4. Estrés: Son las variables que representan el estado de estrés reportado por el sujeto en cuestión, para varios tiempos de referencia. Las etiquetas asociadas a este conjunto de variables son las siguientes:

- *estres_act*: Estado de estrés en que el sujeto en cuestión se encuentra actualmente, de acuerdo con los valores mostrados en la Tabla 2.1.
- *estres1:* Estado de estrés en que el sujeto en cuestión se encontraba hace 1 año.
- *estres5*: Estado de estrés en que el sujeto en cuestión se encontraba hace 5 años.
- *estres10:* Estado de estrés en que el sujeto en cuestión se encontraba hace 10 años.
- *estres20:* Estado de estrés en que el sujeto en cuestión se encontraba hace 20 años.
- *estres30:* Estado de estrés en que el sujeto en cuestión se encontraba hace 30 años.

5. Peso: Son las variables que representan el peso corporal que el sujeto en cuestión reportó tener, para varios tiempos de referencia. Las etiquetas asociadas a este conjunto de variables son las siguientes:
   - *peso_act*: Peso que el sujeto en cuestión reportó tener actualmente, de acuerdo con los valores mostrados en la Tabla 2.1.
   - *peso1:* Peso que el sujeto en cuestión reportó tener hace 1 año.
   - *peso5*: Peso que el sujeto en cuestión reportó tener hace 5 años.
   - *peso10:* Peso que el sujeto en cuestión reportó tener hace 10 años.
   - *peso20:* Peso que el sujeto en cuestión reportó tener hace 20 años.
   - *peso30:* Peso que el sujeto en cuestión reportó tener hace 30 años.

Valor	Significado
1	Muy malo
2	Malo
3	Regular
4	Bueno
5	Muy bueno
6	No sé
7	No aplica
8	No quiero responder

Tabla 2.1. Valores de variables *condi, salud, estres, peso* y su significado.

Variables de "Antropometría"

1. IMC. Índice de masa corporal del sujeto en cuestión. La etiqueta de la base de datos asociada a esta variable es *AIMC*, y su valor asociado está dado a partir de las definiciones de la Tabla 2.2.

AIMC	Valor de IMC
1	<18.5
2	18.5-25
3	25-30
4	30-35
5	35-40
6	>40

Tabla 2.2. Valores de la variable *AIMC*.

Variables de "Datos Personales"

A diferencia de las variables de *Autoevaluación de Salud*, las variables de este conjunto representan datos comprobables y cuyo contenido no depende del criterio del encuestado.

1. Nivel de estudios. Nivel de estudios del sujeto en cuestión. La etiqueta de la base de datos asociada a esta variable es *id_gestud,* y su valor asociado está dado a partir de las definiciones de la Tabla 2.3.

id_gestud	Nivel de estudios
	Ninguno
Prim	Primaria
Sec	Secundaria
Bach	Bachillerato
CarTec	Carrera Técnica
Lic	Licenciatura
Mast	Maestría
Doc	Doctorado
PDoc	Post-Doctorado
Otro	Otro

Tabla 2.3. Valores de la variable *id_gestud*.

2. Puesto. Puesto de trabajo del sujeto en cuestión. La etiqueta de la base de datos asociada a esta variable es *Apuesto*, y su valor asociado está dado a partir de las definiciones de la Tabla 2.4.

Apuesto	Puesto
Acade	Académico
Admin	Administrativo
Asi	Asistente
Coo	Coordinador
E	Estudiante
ED	Estudiante de Doctorado
EM	Estudiante de Maestría
Int	Intendencia
Inv	Investigador
InvE	Investigador Emérito
Jef	Jefe de Área
Lab	Laboratorista
Sec	Secretaria
Tec	Técnico
Vig	Vigilante

Tabla 2.4. Valores de la variable *Apuesto.*

3. Sexo. Sexo del sujeto en cuestión. La etiqueta de la base de datos asociada a esta variable es *id_sexo*, y su valor asociado está dado a partir de las definiciones de la Tabla 2.5.

id_sexo	Sexo
F	Femenino
M	Masculino
O	Otro

Tabla 2.5. Valores de la variable *id_sexo.*

4. Edad: Edad del sujeto en cuestión. La etiqueta de la base de datos asociada a esta variable es $Aedad$, y es una variable continua que representa la edad del sujeto en años.

2.2 – Objetivos del estudio

En términos prácticos, este trabajo es una extensión a más grande escala del estudio referenciado en [1] y, en este sentido, plantea los siguientes objetivos esenciales:

- Describir y analizar cómo el subconjunto de las variables de *Autoevaluación de Salud* descritas anteriormente cambia en el tiempo, con el fin de responder preguntas como:
    - ¿Cuáles variables tienden a preservarse, cuáles tienden a cambiar, y en qué medida?
    - Entendiendo que estas variables están inherentemente relacionadas a *hábitos* de vida, ¿cuál es el tiempo de vida de un hábito?
- Describir cómo son los patrones de evolución de las variables de *Autoevaluación de Salud* en la población en general y en ciertos grupos específicos de personas (definidos a partir de las variables de *Antropometría* y de *Datos Personales*), con el fin de responder preguntas como:
    - ¿Cómo es esta evolución de variables en el tiempo, cuando se analiza a la población en general vs. cuando se analiza a un grupo que comparte una característica específica? ¿Es distinta?
    - ¿Cuáles son los grupos de personas más asociados a ciertos tipos de patrones de evolución? ¿Estos grupos tienen alguna característica conductual que los hace ser más propensos a mostrar este tipo de patrones de evolución de variables?
- Utilizar el conocimiento obtenido a partir de lo anterior para construir diversos modelos predictivos (basados principalmente en el método *Generalized Naive Bayes*, descrito por Stephens et. al en [1]) que permita hacer inferencias respecto al estado de salud actual y futuro de cualquier persona, dado su historial personal de *Variables de Autoevaluación de Salud*.

Capítulo 3. Evolución de hábitos en el tiempo

3.1 – Preprocesamiento de las variables de interés

Como se discutió en el Capítulo 2, tenemos tres subconjuntos de variables de interés: variables de autoevaluación de salud, variables de antropometría y variables de datos personales.

- Variables de autoevaluación de salud

  Descripción general: Son las variables históricas relacionadas a la salud (*Ejercicio, Condición Física, Estrés, Salud* y *Peso*) de la persona en cuestión de los últimos 1, 5, 10, 20 y 30 años. Están directa o indirectamente asociados a hábitos o conductas de vida:
  - El ejercicio es un hábito en sí.
  - El estrés es típicamente causa y efecto del desarrollo de malos hábitos relacionados con la salud (adicciones, falta de sueño, etcétera) [7][8].
  - El peso y la condición física de una persona están asociados directamente al tipo de hábitos de ejercicio y alimentación que dicha persona mantiene.
  - La salud, como un estado general de bienestar físico, mental y social [9], está asociado a una buena dieta, buena actividad física, consumos bajos de alcohol, no fumar, entre otros hábitos [10].

  Importancia y uso dentro de este estudio: Queremos estudiar la evolución de estas variables en el tiempo, aprovechando que la base de datos sobre la que estamos trabajando nos permite reconstruir los historiales de los últimos 10-30 años de cada una de estas variables, para cada persona dentro de la base de datos.

  Etiquetas dentro de la base de datos: *ejer_act, ejer1, ejer5, ejer10, ejer20, ejer30, condi_act, condi1, condi5, condi10, condi20, condi30, salud_act, salud1, salud5, salud10, salud20, salud30, estres_act, estres1, estres5, estres10, estres20, estres30, peso_act, peso1, peso5, peso20* y *peso30.*

- Variables de antropometría y datos personales

  Descripción general: Son las variables que determinan características físicas/sociales actuales (*Índice de Masa Corporal, Nivel de Estudios, Puesto, Sexo y Edad*) de la persona en cuestión.

  Importancia y uso dentro de este estudio: Queremos usar estas variables para dividir a la población general en subgrupos, y estudiar si estos subgrupos muestran

patrones distintos de historiales de salud cuando se comparan contra la población general. En estas variables es donde estamos buscando características conductuales que definan las acciones de las personas en torno a sus hábitos (y, consecuentemente, que definan sus historiales de salud).

Etiquetas dentro de la base de datos: *AIMC, id_gestud, Apuesto, id_sexo, Aedad.*

Todas estas variables de interés son continuas (para el caso de las variables de *Ejercicio y Aedad*) o categóricas (todas las demás). Para facilitar los análisis que se realizarán más adelante, debemos realizar un preprocesamiento de estas variables y aplicar una *clusterización* sobre ellas, de manera que todas puedan analizarse como si fueran variables categóricas con 2, 3 o 4 categorías distintas máximo. El razonamiento tras esta necesidad, así como la metodología seguida para aplicar esta clusterización se explica en las subsecciones siguientes.

3.1.1 – Preprocesamiento de las variables de *Autoevaluación de Salud*

Antes del procesamiento, las variables asociadas a *Ejercicio* son variables continuas que toman valores superiores mayores que 0, y las variables asociadas a *Condición Física, Salud, Estrés* y *Peso* son variables categóricas que pueden tomar todas 8 valores distintos, según lo descrito en la Tabla 2.1 del Capítulo 2.1:

Valor	Significado
1	Muy malo
2	Malo
3	Regular
4	Bueno
5	Muy bueno
6	No sé
7	No aplica
8	No quiero responder

Como se mencionó anteriormente, queremos estudiar la evolución de estas variables en el tiempo a través de la reconstrucción de los historiales de cada persona dentro de la base de datos. Es decir, para cada persona y cada conjunto de etiquetas asociadas a una variable de *Autoevaluación de Salud*, queremos reconstruir una historia como la mostrada en la Figura 3.1.

condi_ act	*condi1*	*condi5*	*condi10*	*condi20*	*condi30*
[1, 8]	[1, 8]	[1, 8]	[1, 8]	[1, 8]	[1, 8]

Figura 3.1. Ejemplo de reconstrucción del historial de *Condición Física* para una persona dentro de la base de datos, partiendo del hecho que cada variable *condiX* puede tomar 8 valores distintos, según la Tabla 2.1.

No obstante, por ejemplo, si intentamos reconstruir el historial de las 6 variables asociadas a *Condición Física* de una persona (*condi_ act, condi1, condi5, condi10, condi20* y *condi30*), usando las 8 categorías mostradas en la Tabla 2.1, este historial tomaría $6^8 = 1{,}679{,}616$ posibles configuraciones. Incluso asumiendo que la gran mayoría de las personas contestó con valores del 1-5, descartando así las categorías 6, 7 y 8, nuestros historiales podrían tomar $6^5 = 7{,}776$ configuraciones posibles. Esto es un problema, pues nuestra base de datos sólo cuenta con 1076 entradas y queremos evitar al máximo reconstruir historiales "únicos". Para el caso de las variables asociadas a *Ejercicio*, al ser variables continuas (pues representan el número de horas de ejercicio semanales reportadas por la persona encuestada), estás deben ser categorizadas antes de poder ser reconstruidas como un "historial".

Con el fin de atender este problema, y generalizar la estructura de los historiales que se reconstruirán para cada persona, hacemos la siguiente clusterización de variables:

- Condición Física. Sea *condiX* la variable asociada a *Condición Física* en el tiempo de referencia "*X*", donde el símbolo $X \in \{act, 1, 5, 10, 20, 30\}$, definimos una nueva variable *condiXB*, que representa a la variable *condiX* clusterizada bajo la siguiente regla:

$$condiXB = \begin{cases} A, si\ condiX \in \{4, 5\} \\ B, si\ condiX\ \in \{1, 2, 3\} \\ N, en\ otro\ caso \end{cases}$$

  Cualitativamente hablando, si un individuo de la base de datos tiene $condiXB = A$, decimos que este individuo reportó una condición física "buena" para el tiempo de referencia representado por el símbolo *X*. Conversamente, si para ese individuo tenemos que $condiXB = B$, decimos que este individuo reportó una condición física "mala" para el tiempo de referencia *X*.

  Siguiendo el mismo razonamiento planteado para *Condición Física,* hacemos las definiciones de *Salud, Estrés* y *Peso*.

- Salud. Definimos una nueva variable *saludXB*, que representa a la variable *saludX* clusterizada bajo la siguiente regla:

$$saludXB = \begin{cases} A, si\ saludX \in \{4,5\} \\ B, si\ saludX \in \{1,2,3\} \\ N, en\ otro\ caso \end{cases}$$

- Estrés. Definimos una nueva variable *estresXB*, que representa a la variable *estresX* clusterizada bajo la siguiente regla:

$$estresXB = \begin{cases} A, si\ estresX \in \{4,5\} \\ B, si\ estresX \in \{1,2,3\} \\ N, en\ otro\ caso \end{cases}$$

- Peso. Definimos una nueva variable *pesoXB*, que representa a la variable *pesoX* clusterizada bajo la siguiente regla:

$$pesoXB = \begin{cases} A, si\ pesoX \in \{3,4,5\} \\ B, si\ pesoX \in \{1,2\} \\ N, en\ otro\ caso \end{cases}$$

  La razón tras usar una definición distinta de "A" en *Peso*, respecto a las demás variables de *Autoevaluación de Salud*, radica en el hecho de que, usando la definición donde "A" representa a las categorías 4 y 5, y "B" representa a las categorías 1, 2, y 3, las clases "A" y "B" resultaban estar muy desbalanceadas: la gran mayoría de la población se encontraba en la clase "B". Esto se debe a que la mayoría de las personas que consideran su peso como "Regular", en realidad no tienen obesidad (simplemente no tienen el peso que desearían tener). Por ello, con el fin de mantener cierta consistencia con respecto a lo que este trabajo considera como "malo" cuando se habla del peso corporal de una persona, se tomó la decisión de ubicar a todas aquellas personas que consideran su peso corporal como "regular" (i.e. $pesoX = 3$) dentro de la categoría "A" de *pesoXB*.

- Ejercicio. Definimos una nueva variable *ejerXB,* que representa a la variable *ejerX* clusterizada bajo la siguiente regla:

$$ejerXB = \begin{cases} A, si\ ejerX > 2.5 \\ B, si\ ejerX < 2.5 \\ N, en\ otro\ caso \end{cases}$$

  La razón tras elegir esta regla de categorización es que el ejercicio semanal recomendado por la *World Health Organization (WHO)* es de al menos 2.5 horas [11]. Así pues, usando esta categorización, se mantiene la misma convención usada para las otras variables donde si un individuo tiene $ejerXB = A$, decimos que este

individuo reportó una cantidad de ejercicio "buena" para el tiempo de referencia representado por el símbolo X. Conversamente, decimos que la cantidad de ejercicio que el individuo reportó es "mala" si $ejerXB = B$.

Utilizando las categorizaciones anteriores, la reconstrucción del historial de la Figura 3.1, pasa a ser como el mostrado en la Figura 3.2.

condi_ actB	*condi1B*	*condi5B*	*condi10B*	*condi20B*	*condi30B*
{A, B, N}	{A, B, N}	{A, B, N}	{A, B, N}	{A, B, N}	{A, B, N}

Figura 3.2. Ejemplo de reconstrucción del historial de *Condición Física* para una persona dentro de la base de datos, utilizando las definiciones *condiXB*.

3.1.2 – Preprocesamiento de las variables de *Antropometría y Datos Personales*

Antes del preprocesamiento, las variables *AIMC, id_ gestud, Apuesto, id_ sexo* y *Aedad* son todas categóricas, según las definiciones planteadas en el Capítulo 2.1. Como se mencionó al inicio de este capítulo, queremos "utilizar estas variables para dividir a la población general en subgrupos y estudiar si estos subgrupos muestran patrones distintos de historiales de salud cuando se comparan contra la población general". Con el fin de reducir la cantidad de subgrupos de la población que estas variables describen, definiremos una serie de nuevas "clases" binarias, establecidas a partir de las variables de referencia. La pertenencia o no pertenencia de una persona a estas nuevas clases debería poder ayudarnos a "perfilar" a dicha persona (i.e. a dar una descripción cualitativa de esta persona) y debería poder ayudarnos a explicar las diferencias que existen entre sus historiales de salud vs. el de la población en general, en caso de que las haya. Así pues, a partir de estas variables, definimos 4 nuevas clases (variables) binarias: la clase *Obesidad,* la clase *Alto grado de estudios*, la clase *Académico* y la clase *Sexo*, mismas que se describen a continuación.

- Obesidad – Clase "O"

  Representa a los individuos de la base de datos que tienen obesidad según la definición de la *WHO*, que establece que un individuo tiene obesidad si su índice de masa corporal es mayor o igual que 4 [12]. Partiendo de la variable *AIMC*, podemos pues definir a la clase "Obesidad" (a la que nos referiremos también como *Clase O*) de la siguiente manera:

$$Obesidad = \begin{cases} 0, si\ AIMC < 4 \\ 1, si\ AIMC \geq 4 \end{cases}$$

La distribución de la población según su pertenencia o no pertenencia a la *Clase O* se puede apreciar en la Figura 3.3.

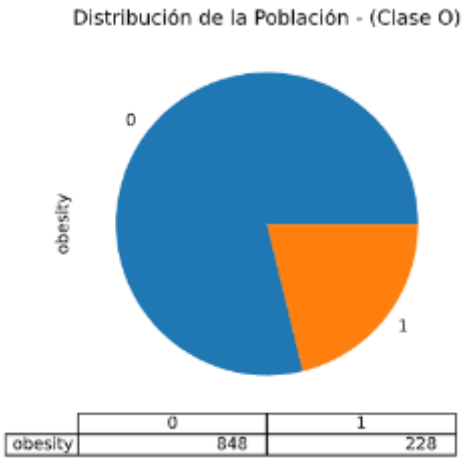


Figura 3.3. Distribución de la población, según la *Clase O*.

- Alto grado de estudios – Clase "H"

Representa a los individuos de la base de datos cuyo grado de estudios es de Licenciatura o superior. Partiendo de la variable *id_ gestud* y los valores mostrados en la Tabla 2.3 (Capítulo 2), podemos pues definir a la clase "Alto grado de estudios" (a la que nos referiremos también como *Clase H*) de la siguiente manera:

$$Alto\ Grado\ de\ Estudios = \begin{cases} 0, si\ id\_gestud\ \in \{Prim, Sec, Bach, CarTec, Otro\} \\ 1, si\ id\_gestud\ \in \{Lic, Mast, Doc, PDoc\} \end{cases}$$

La distribución de la población según su pertenencia o no pertenencia a la *Clase H* se puede apreciar en la Figura 3.3.

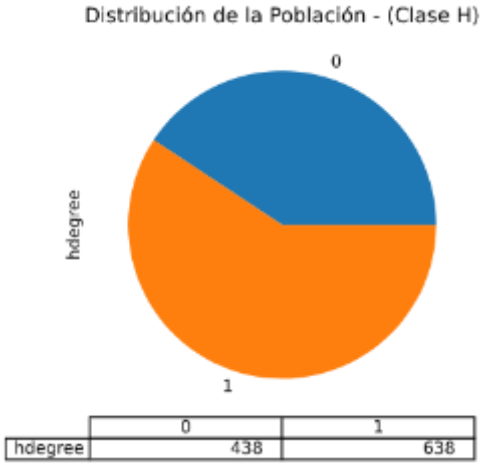


Figura 3.4. Distribución de la población, según la *Clase H*.

- Académico – Clase "A"

Representa a los individuos de la base de datos cuya ocupación/puesto de trabajo es del tipo académico (i.e. académicos e investigadores). Partiendo de la variable *Apuesto* y los valores mostrados en la Tabla 2.4 (Capítulo 2), podemos pues definir a la clase "Académico" (a la que nos refereriemos también como la *Clase A*) de la siguiente manera:

$$Académico = \begin{cases} 0, si\ Apuesto\ \in \{Admin, Asi, Coo, E, ED, EM, Int, Jef, Lab, Sec, Tec, Vig\} \\ 1, si\ Apuesto\ \in \{Acade, Inv, InvE\} \end{cases}$$

La distribución de la población según su pertenencia o no pertenencia a la *Clase A* se puede apreciar en la Figura 3.5.

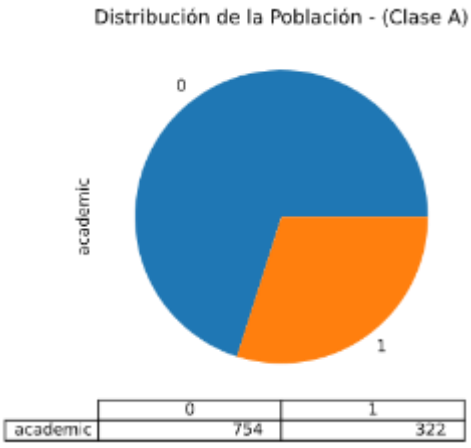


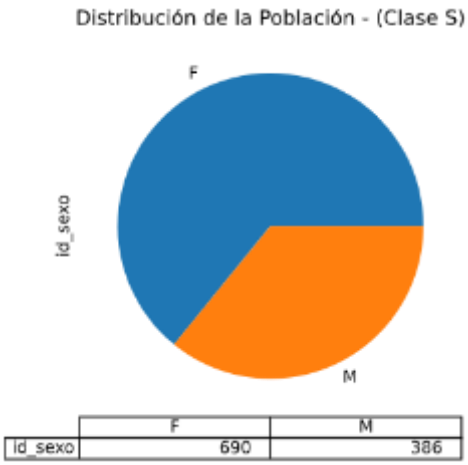
Figura 3.5. Distribución de la población, según la *Clase A*.

- Sexo – Clase "S"

Diferencia a la población entre mujeres y hombres. Esta clase está tomada directamente de la variable de la base de datos *id_ gestud*, y sólo se renombró para ser consistente con las otras clases de interés que se van a analizar. Esta clase se define de la siguiente manera:

$$Sexo = \begin{cases} \text{F}, si\ la\ persona\ en\ cuestión\ es\ mujer \\ M, si\ la\ persona\ en\ cuestión\ es\ hombre \end{cases}$$

La distribución de la población según su símbolo asociado (“F” o “M”) se puede apreciar en la figura 3.6



	F	M
id_sexo	690	386

Figura 3.6. Distribución de la población, según la *Clase S*.

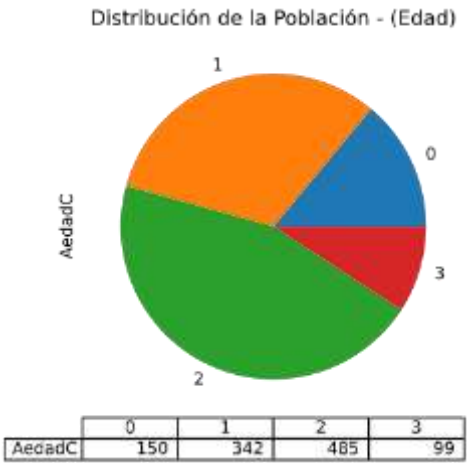
Además de las nuevas variables binarias anteriores, definimos una nueva variable cuyo propósito sea el de convertir la variable $Aedad$ en una variable categórica. Así pues:

- Edad

  Representa el grupo de edad en el que se encuentran los individuos de la base de datos. Partiendo de la variable $Aedad$, podemos definir a la clase “Edad” de la siguiente manera:

$$Edad = \begin{cases} 0, si\ 15 \leq Aedad \leq 28 \\ 1, si\ 28 < Aedad \leq 40 \\ 2, si\ 40 < Aedad \leq 60 \\ 3, si\ 60 < Aedad \leq 90 \end{cases}$$

  La distribución de la población según la variable $Edad$ se puede apreciar en la Figura 3.7.



	0	1	2	3
AedadC	150	342	485	99

Figura 3.7. Distribución de la población, según la variable *Edad*.

3.1.3 – Discusión de la disponibilidad de información de las variables en la base datos

Los datos correspondientes a las variables presentadas en la sección 3.1.1 no están completos para todas las entradas de la base de datos. Es decir, hay entradas para las cuales la persona encuestada no dio una respuesta, o bien, para las cuales no tienen información disponible (por ejemplo, para el caso de variables históricas de hace 20 o 30 años, en personas menores a 30 años). En general, hay al menos 930 entradas de la base de datos que tienen información completa para cada variable de *Autoevaluación de Salud* en el rango de 0 – 20 años, y hay al menos 650 entradas que tienen información completa para el rango completo de 0 – 30 años. Para el rango de 0 – 10 años, hay al menos 1060 entradas disponibles para cada variable.

Los datos correspondientes a las variables presentadas en la sección 3.1.2 sí están completos para todas las entradas de la base de datos.

3.2 – Análisis de la evolución de las variables de *Autoevaluación de Salud* en el tiempo

En primer lugar, queremos cuantificar y describir cómo los subconjuntos de las variables de *Autoevaluación de Salud* descritas en la sección 3.1.1 cambian en el tiempo, tanto en la población general como en poblaciones específicas. Esto con el fin de entender cuáles variables tienden a preservarse, cuáles tienden a cambiar y en qué medida. Para hacer esto, para cada variable de *Autoevaluación de Salud* y diversas segmentaciones de la población, haremos un "análisis de supervivencia de hábitos", cuya metodología se explica a continuación.

3.2.1 – Metodología del "Análisis de supervivencia de hábitos"

Como ya se ha mencionado anteriormente, partimos de la hipótesis de que las variables de *Autoevaluación de Salud* están directa o indirectamente relacionadas a hábitos. Por esta razón, a partir de ahora, hablaremos de "hábitos" como sinónimo de la categoría que puede tomar cada una de estas variables ("A", "B" o "N", según las clusterizaciones hechas en la sección 3.1.1). Es decir, si una persona reporta una categoría "A" para cierta variable, diremos que esa persona reportó un hábito "A" para esa variable.

Partiendo del hecho de que todas nuestras variables de *Autoevaluación de Salud* toman sólo tres valores: "A", "B" y "N", y que sólo los valores de "A" y "B" se consideran como "evidencia real" (pues la categoría "N" fue usada sólo para aquellas entradas para las cuales la persona encuestada no hubiera reportado ninguna respuesta), tenemos cuatro tipos distintos de patrones de evolución de hábitos.

1. (Patrón A -> A). Decimos que tenemos un "patrón de preservación de hábitos *buenos* con relación a la variable a en el tiempo t", si una persona que actualmente reporta una condición de hábitos "buena" de esa variable (i.e. $a_{act} = A$), para la misma variable hace t años también reportó una condición de hábitos "buena" (i.e. $a_t = A$). Por ejemplo, una persona muestra un patrón de preservación de hábitos bueno con relación a su *Condición Física* en los últimos 10 años si, para esa persona, en la base de datos leemos que $condi\_actB = A$ y $condi10B = A$.
2. (Patrón B -> B). Decimos que tenemos un "patrón de preservación de hábitos *malos* con relación a la variable a en el tiempo t", si una persona que actualmente reporta una condición de hábitos "mala" de esa variable (i.e. $a_{act} = B$), para la misma variable hace t años también reportó una condición de hábitos "mala" (i.e. $a_t = B$). Por ejemplo, una persona muestra un patrón de preservación de hábitos malo con relación a su *Condición Física* en los últimos 10 años, si para esa persona, en la base de datos leemos que $condi\_actB = B$ y $condi10B = B$.
3. (Patrón B -> A). Siguiendo el mismo razonamiento, tenemos un "patrón de cambio de hábitos *malo a bueno* con relación a la variable a en el tiempo t si $a_{act} = A$ y $a_t = B$.
4. (Patrón A -> B). Por último, tenemos un "patrón de cambio de hábitos *bueno a malo* con relación a la variable a en el tiempo t si $a_{act} = B$ y $a_t = A$.

En términos generales, el "análisis de supervivencia de hábitos" busca cuantificar las "fuerzas" de preservación y de cambio de hábitos con relación a las variables de *Autoevaluación de Salud*, para varios tiempos de referencia, asociadas a estos cuatro tipos de patrones. Es decir, para cualquier variable de *Autoevaluación de Salud* "a" este análisis pretende calcular:

1. La fuerza del patrón *bueno-bueno* de preservación de hábitos con relación a la variable a.
2. La fuerza del patrón *malo-malo* de preservación de hábitos con relación a la variable a.
3. La fuerza del patrón *malo-bueno* de cambio de hábitos con relación a la variable a.
4. La fuerza del patrón *bueno-malo* de cambio de hábitos con relación a la variable a.

Queremos expresar a las cuatro fuerzas anteriores como una proporción de las personas que, en un determinado tiempo de referencia, están preservando o cambiando sus hábitos. Para hacer esto, considérese el siguiente mecanismo:

- Sea P la cantidad de personas que hace t años reportaron un hábito X para la variable de *Autoevaluación de Salud* a (i.e. que reportaron $a_t = X$), y sea P' la cantidad de personas que, además de haber reportado el mismo hábito X hace t años, actualmente reporta un hábito Y para esa misma variable (i.e. P' es el número de personas que reportaron $a_t = X$ y $a_{act} = Y$), queremos expresar a la fuerza de preservación/cambio del hábito X al hábito Y de la variable a, es decir $F^t_{X\to Y}(a)$, como:

$$F^t_{X\to Y}(a) = \frac{P'}{P}$$

$$F^t_{X\to Y}(a) = \frac{N_{a_{act}=Y,\, a_t=X}}{N_{a_t=X}}$$

Es fácil ver que la función propuesta no es más que una probabilidad condicional, pues:

$$P(a_{act} = Y | a_t = X) = \frac{P(a_{act} = Y \cap a_t = X)}{P(a_t = X)} = \frac{\frac{N_{a_{act}=Y,\, a_t=X}}{N}}{\frac{N_{a_t=X}}{N}} = \frac{N_{a_{act}=Y,\, a_t=X}}{N_{a_t=X}}$$

Por lo tanto:

$$F^t_{X\to Y}(a) = P(a_{act} = Y | a_t = X)$$

Esta interpretación de las "fuerzas de preservación/cambio de hábitos", como una medida de probabilidad condicional tiene sentido, pues nos estamos basando en la noción de cuantificar la probabilidad de un evento (i.e. que una persona reporte un cierto tipo de hábito en el tiempo actual) dada una evidencia (i.e. que esa persona haya reportado un cierto tipo de hábito en el tiempo de referencia pasado).

Partiendo de la definición anterior, proponemos pues la siguiente serie de funciones de fuerzas de preservación y de cambio de hábitos, con la intención de cuantificar la evolución de los cuatro patrones de interés.

1. Fuerza de preservación de hábitos *buenos* en el rango de tiempo t ($F^t_{A\to A}$).

$$F^t_{A\to A}(a) = P(a_{act} = A | a_t = A)$$

2. Fuerza de preservación de hábitos *malos* en el rango de tiempo t ($F^t_{B\to B}$).

$$F^t_{B\to B}(a) = P(a_{act} = B | a_t = B)$$

3. Fuerza de cambio de hábitos *malo a bueno* en el rango de tiempo t ($F^t_{B\to A}$).

$$F^t_{B\to A}(a) = P(a_{act} = A | a_t = B)$$

4. Fuerza de cambio de hábitos *bueno a malo* en el rango de tiempo t ($F^t_{A\to B}$).

$$F^t_{A\to B}(a) = P(a_{act} = B | a_t = A)$$

El "análisis de supervivencia de hábitos" consistirá pues en el calcular estas cuatro fuerzas de preservación/cambio de hábitos, para cada variable de *Autoevaluación de Salud*, usando todos los tiempos de referencia disponibles (0, 1, 5, 10, 20 y 30 años), para toda la población y grupos específicos de ésta (según las clases *O, H, A* y *S* definidas en la sección 3.1.2).

Por ejemplo, para determinar las fuerzas de preservación/cambio de hábitos en torno al historial de *Ejercicio*, partimos de los siguientes conteos hechos a la base de datos.

Tiempo de referencia	$N_{ejer_t=A}$	$N_{ejer_t=B}$
$t = 1$	465	608
$t = 5$	565	509
$t = 10$	593	475
$t = 20$	500	440
$t = 30$	317	341

Tiempo de referencia	$N_{ejer_{act}=A,ejer_t=A}$	$N_{ejer_{act}=A,ejer_t=B}$	$N_{ejer_{act}=B,ejer_t=A}$	$N_{ejer_{act}=B,ejer_t=B}$
$t = 1$	324	99	141	509
$t = 5$	306	117	259	392
$t = 10$	285	136	308	339
$t = 20$	224	137	276	303
$t = 30$	133	104	184	237

Donde $N_{ejer_t=X}$ es el número de personas dentro de la base de datos que reportaron un hábito de tipo X en el tiempo t, y $N_{ejer_{act}=Y, ejer_t=Z}$ es el número de personas que reportó un hábito de ejercicio Y en el tiempo actual, y un hábito de ejercicio Z en el tiempo t. Así pues, para calcular las fuerzas de preservación/cambio de hábitos de *Ejercicio* $F^t_{A\to A}$, $F^t_{B\to B}$, $F^t_{B\to A}$ y $F^t_{A\to B}$ en distintos intervalos de t, usamos la definición presentada anteriormente. Es decir:

$$F^t_{X\to Y}(Ejercicio) = P(ejer_{act} = Y | ejer_t = X):$$

Obteniendo así:

	$F^t_{A\to A}$	$F^t_{B\to B}$,	$F^t_{B\to A}$	$F^t_{A\to B}$
$t = 1$	0.697	0.837	0.163	0.303
$t = 5$	0.542	0.77	0.23	0.458
$t = 10$	0.481	0.714	0.286	0.519
$t = 20$	0.448	0.689	0.311	0.552
$t = 30$	0.42	0.695	0.305	0.58

Con el fin de ilustrar gráficamente las tendencias en estas evoluciones de hábitos, los resultados de este análisis se muestran en el gráfico de la figura Figura 3.8. Todos los análisis de este tipo hechos para las demás variables de autoevaluación de salud (y demás subgrupos de la población) se presentarán, a partir de este punto, a través de un gráfico similar. Las curvas del gráfico están coloreadas por tipo de patrón de evolución de hábitos

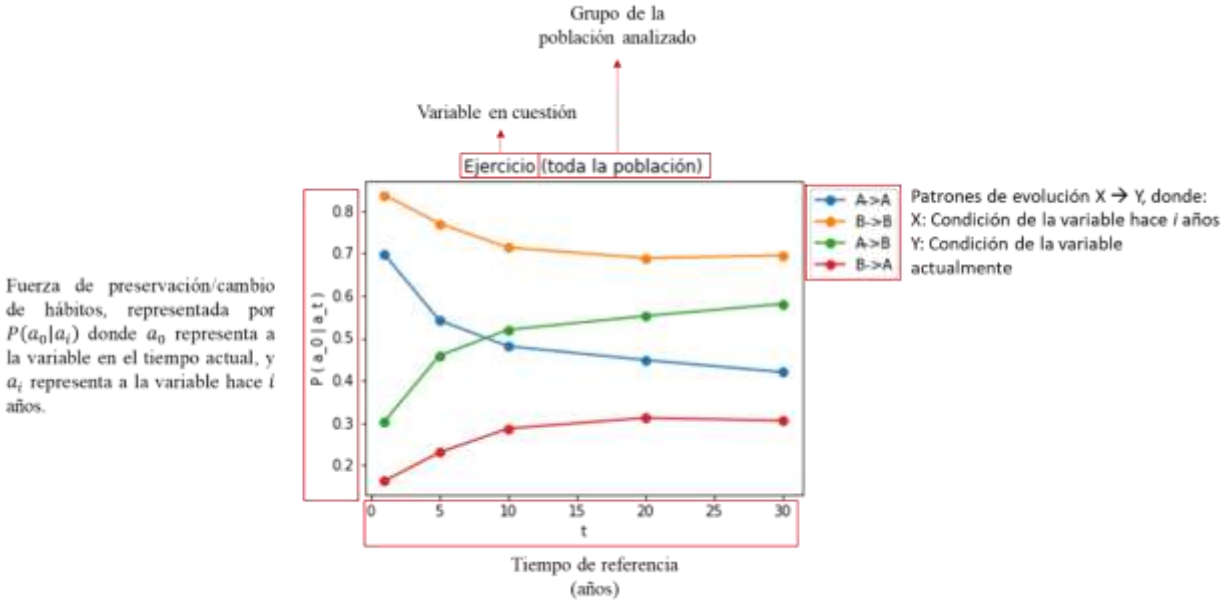


Figura 3.8. Desglose del gráfico de la fuerza de preservación/cambio de hábitos aplicado a toda la población del estudio.

Las fuerzas de preservación/cambio de hábitos sólo nos dan valores en el rango [0, 1] y ellas no nos dan una idea clara de la cantidad de personas que está preservando/cambiando sus hábitos en términos absolutos. Para reflejar estas cantidades absolutas de personas que están preservando/cambiando sus hábitos en cada "tiempo de muestreo" de las variables de *Autoevaluación de Salud* (0, 1, 5, 10, 15 y 20 años) con respecto al tiempo de referencia inmediato anterior, usamos *funnels* como el que se muestra en la Figura 3.9. En cada etapa se muestra la cantidad de personas que *sigue* manteniendo un hábito, con respecto de los tiempos de referencia anteriores. Con la información disponible, podrían hacerse *funnels* de 2, 3, 4, 5 y 6 etapas. No obstante, ya que, como se discutió en la sección 3.1.3, no hay datos de las variables de *Autoevaluación de Salud* completos para todas las entradas de la base de datos, se decidió mostrar *funnels* de sólo 5 etapas, pues para los tiempos de referencia que esas etapas engloban (0 – 20 años), hay al menos 930 entradas disponibles completas para cada variable de interés.

Si bien el *funnel* refleja gráficamente sólo a la cantidad de personas que presentan patrones de preservación de hábitos (buenos o malos), las cantidades de las personas que presentan patrones de cambio de hábitos se pueden inferir a partir del mismo gráfico (i.e. las personas que cambian su hábito son aquellas que no lo mantienen).

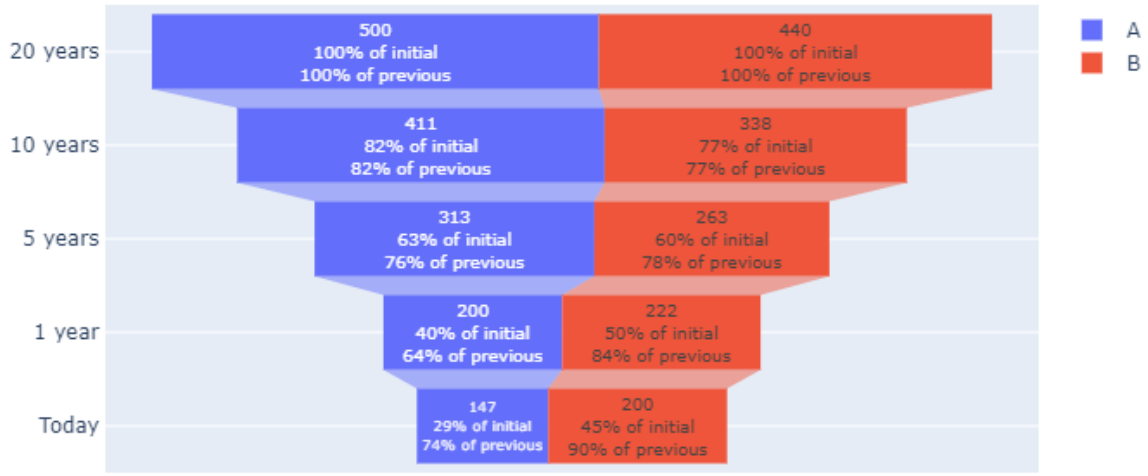


Figura 3.9. *Funnel* de cinco etapas (20 años, 10 años, 5 años, 1 año y actual) de preservación de hábitos.

3.2.2 – Resultados del "Análisis de supervivencia de hábitos"

Los resultados de los análisis de supervivencia de hábitos para las variables de *Autoevaluación de Salud* se muestran en las subsecciones siguientes.

3.2.2.1 – *Ejercicio*

Al analizar el comportamiento de la población general (Figura 3.10), observamos que la fuerza de preservación de hábitos disminuye cuando aumenta el tiempo de referencia, es decir, es más factible preservar hábitos de ejercicio (buenos o malos) en el corto plazo ($<$ 5 años) que en el mediano o largo plazo ($>$ 5 años). Asimismo, la fuerza de cambio de hábitos aumenta cuando aumenta el tiempo de referencia, es decir, es más factible cambiar hábitos de ejercicio en el largo plazo que en el corto.

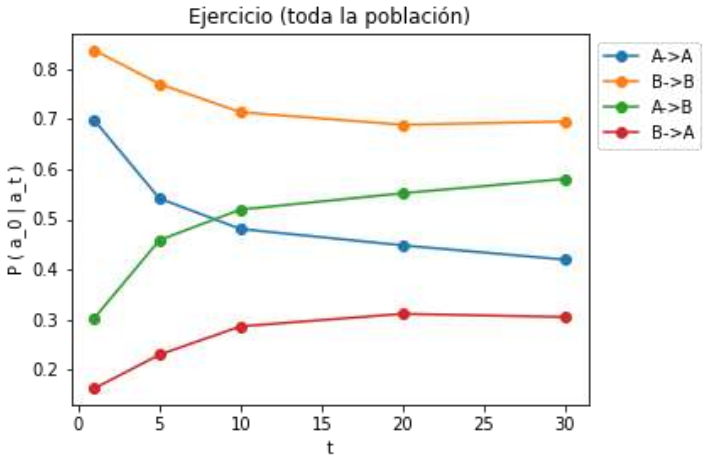


Figura 3.10. Gráfico de las fuerzas de preservación/cambio de hábitos de la variable *Ejercicio,* correspondientes a la población general a lo largo de 30 años.

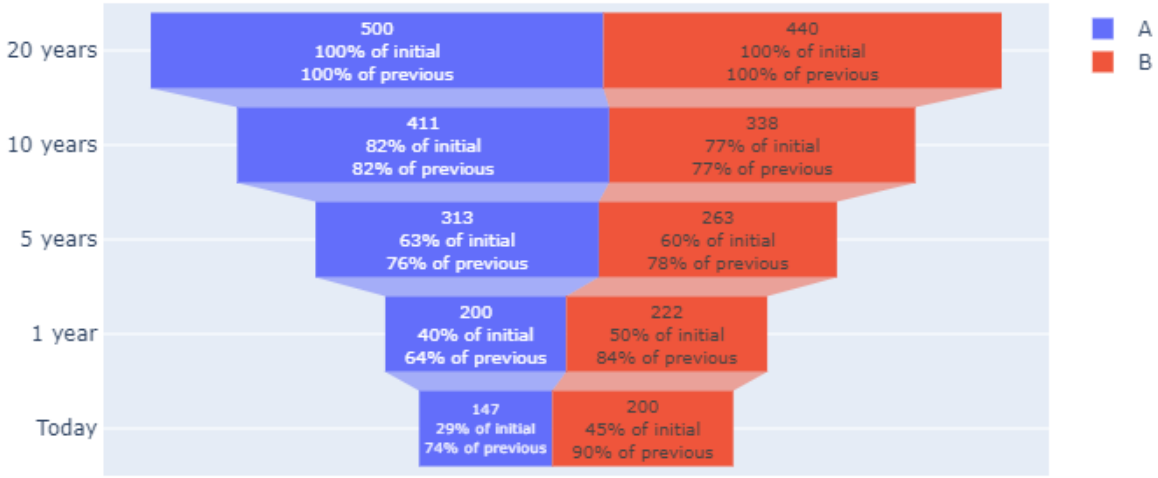


Figura 3.11. Funnel de cinco etapas de preservación de hábitos de *Ejercicio* correspondientes a la población general.

Al dividir a la población en subgrupos, se observan claras diferencias entre distintos subgrupos; por un lado, tal y como se muestra en las figuras 3.12 – 3.15, las personas con obesidad (clase *O*) mostraron ser, históricamente, considerablemente más propensas a mantener un mal hábito de ejercicio y menos propensos a mantener un buen hábito de ejercicio con respecto de las personas sin obesidad (clase *NO*). Del mismo modo, las personas pertenecientes a la clase *O* muestran a ser más propensas a cambiar un buen hábito de ejercicio por uno malo, y menos propensas a cambiar un mal hábito por uno bueno, que las personas de la clase *NO*.

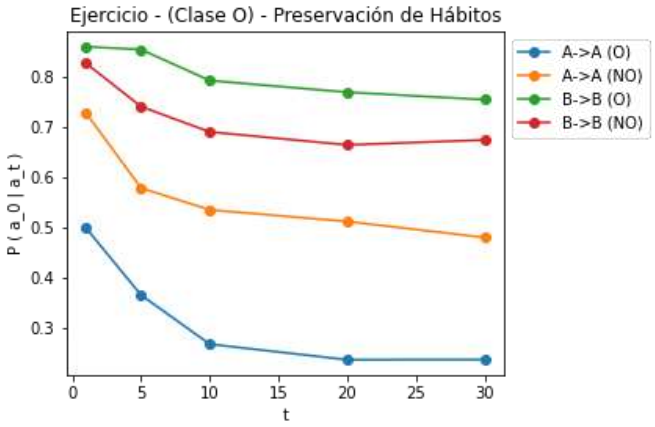


Figura 3.12. Gráfico de las fuerzas de preservación de hábitos de la variable *Ejercicio,* correspondientes a la división de la población *O vs. NO*.

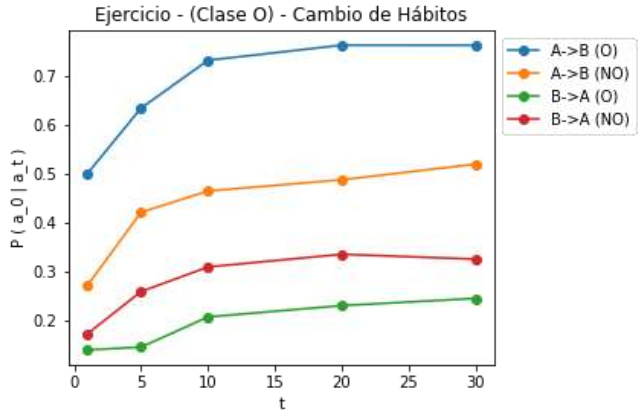


Figura 3.13. Gráfico de las fuerzas de cambio de hábitos de la variable *Ejercicio,* correspondientes a la división de la población *O vs. NO*.

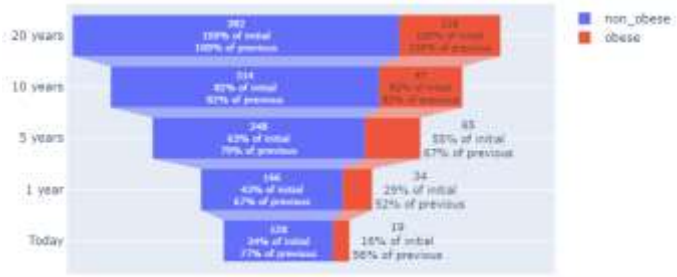


Figura 3.14. *Funnel* de cinco etapas de preservación de hábitos buenos de *Ejercicio*, correspondiente a la división de la población *O* vs. *NO*.

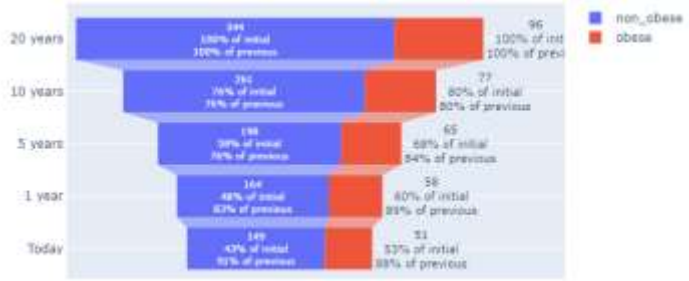


Figura 3.15. *Funnel* de cinco etapas de preservación de hábitos malos de *Ejercicio*, correspondiente a la división de la población *O* vs. *NO*.

Por su parte, tal como se observa en las figuras 3.16-3.19, las personas con alto grado de estudios (clase *H*) mostraron ser históricamente más propensas a mantener un buen hábito de ejercicio y menos propensas a mantener uno malo que las personas sin alto grado de estudios (clase *NH*). Del mismo modo, éstas (clase *H*) muestran ser menos propensas a cambiar un buen hábito de ejercicio por uno malo, y más propensas a cambiar un mal hábito por uno bueno que las personas de la clase *NH*. Por sí solo, esto implica que las

personas con alto grado de estudios son aparentemente más conscientes de los beneficios que trae consigo el mantener buenos hábitos de ejercicio.

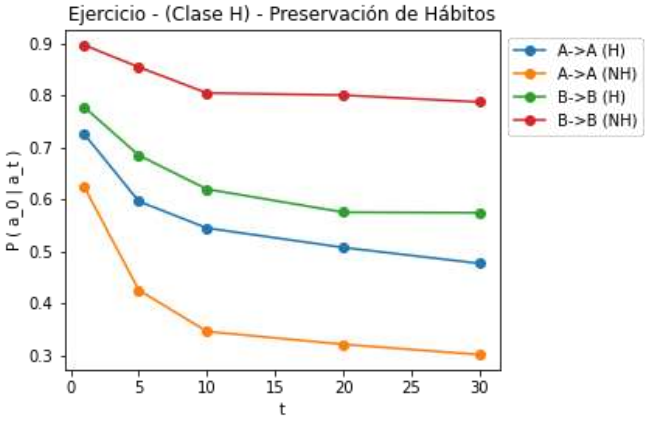


Figura 3.16. Gráfico de las fuerzas de preservación de hábitos de la variable *Ejercicio,* correspondientes a la división de la población *H vs. NH.*

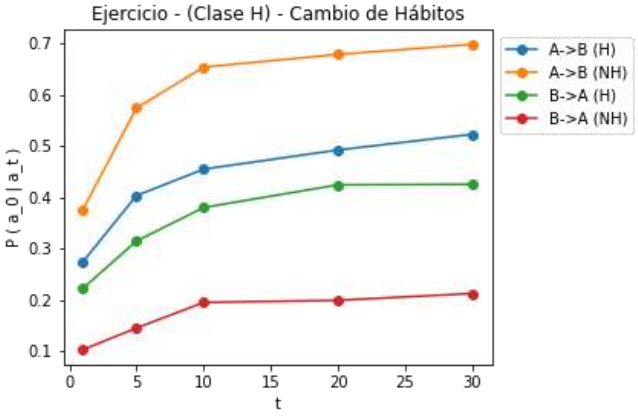


Figura 3.17. Gráfico de las fuerzas de cambio de hábitos de la variable *Ejercicio,* correspondientes a la división de la población *H vs. NH.*

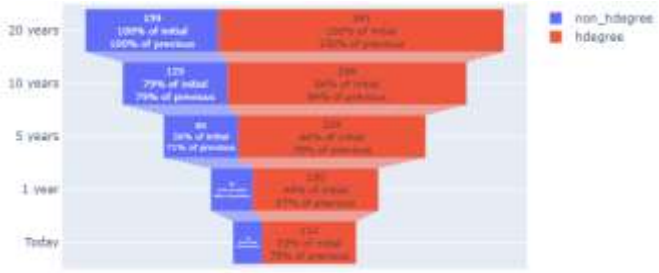


Figura 3.18. *Funnel* de cinco etapas de preservación de hábitos buenos de *Ejercicio*, correspondiente a la división de la población *H* vs. *NH.*

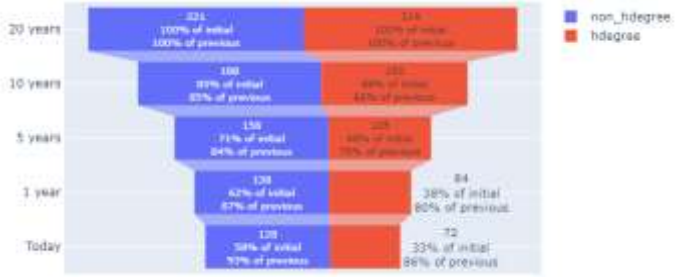


Figura 3.19. *Funnel* de cinco etapas de preservación de hábitos malos de *Ejercicio*, correspondiente a la división de la población *H* vs. *NH.*

Por otro lado, analizando a la división dada por la clase A, puede observarse que las personas con puestos académico (clase *A*) mostraron ser históricamente más propensas a mantener un buen hábito de ejercicio, pero mostraron, al mismo tiempo tener patrones similares de manutención de malos hábitos de ejercicio con respecto a las personas de la clase *NA* en los últimos 10 años. Este comportamiento se puede observar claramente en las figuras 3.20 y 3.21 (las líneas color verde y roja representan a los académicos y no académicos, respectivamente). Así pues, en el contexto de la variable *Ejercicio*, la única diferencia de comportamiento que los académicos muestran contra la población no académica parece ser que los académicos abandonan un hábito de ejercicio que ya es bueno con menor frecuencia que los no académicos. Éste es un fenómeno único dentro del análisis de la variable *Ejercicio* y habla de que la gente académica no necesariamente persigue buenos hábitos sólo porque éstos sean buenos (pues el nivel de cambio malo -> bueno no

es muy diferente contra la población de no académicos), sino que simplemente las personas académicas tienen más constancia en la preservación de sus hábitos en general (sean buenos o malos).

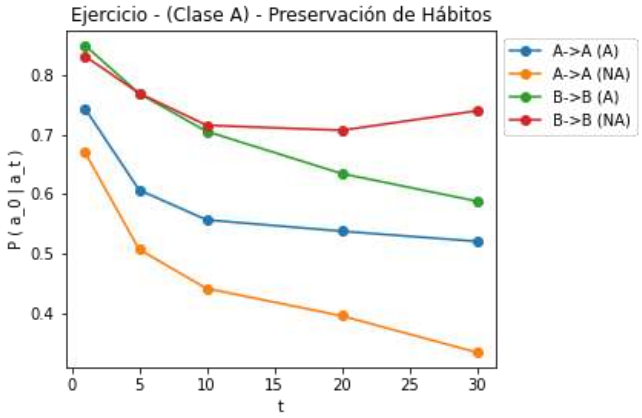


Figura 3.20. Gráfico de las fuerzas de preservación de hábitos de la variable *Ejercicio,* correspondientes a la división de la población *A vs. NA*.

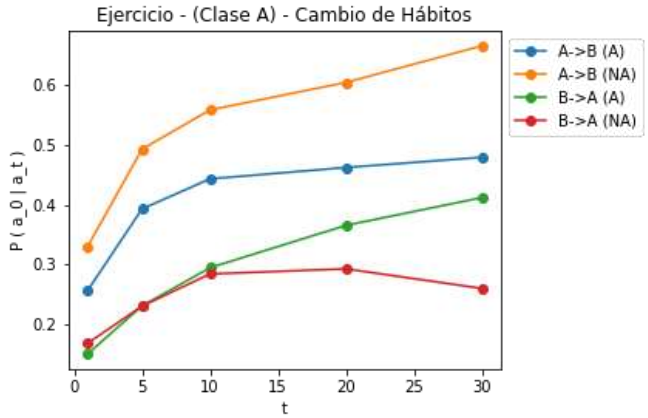


Figura 3.21. Gráfico de las fuerzas de cambio de hábitos de la variable *Ejercicio,* correspondientes a la división de la población *A vs. NA*.

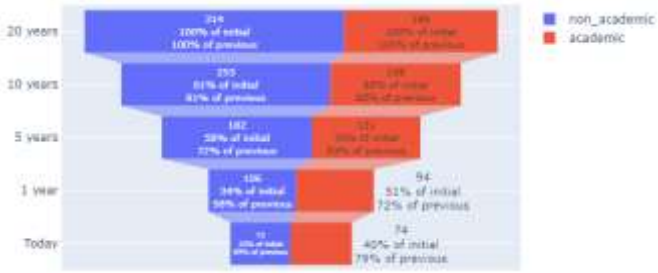


Figura 3.22. *Funnel* de cinco etapas de preservación de hábitos buenos de *Ejercicio*, correspondiente a la división de la población *A* vs. *NA*.

Figura 3.23. *Funnel* de cinco etapas de preservación de hábitos malos de *Ejercicio*, correspondiente a la división de la población *A* vs. *NA*.

Al respecto de la división de la clase S, es de notar que no parece haber diferencia significativa entre los historiales de la variable *Ejercicio* entre hombres y mujeres. La similitud entre las tendencias y generalidades de ambas se observan con más claridad en las figuras 3.24 y 3.25.

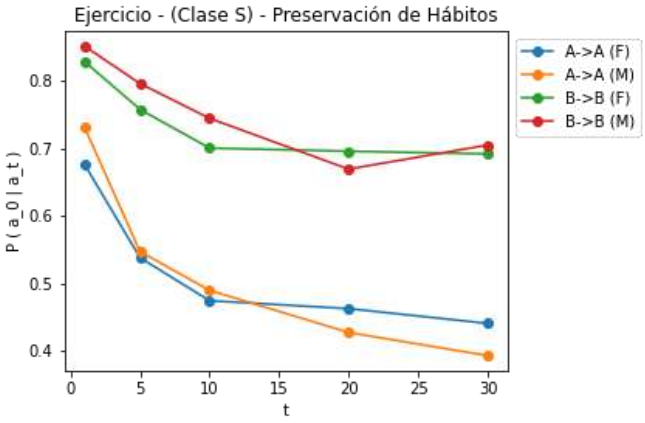


Figura 3.24. Gráfico de las fuerzas de preservación de hábitos de la variable *Ejercicio,* correspondientes a la división de la población *F vs. M.*

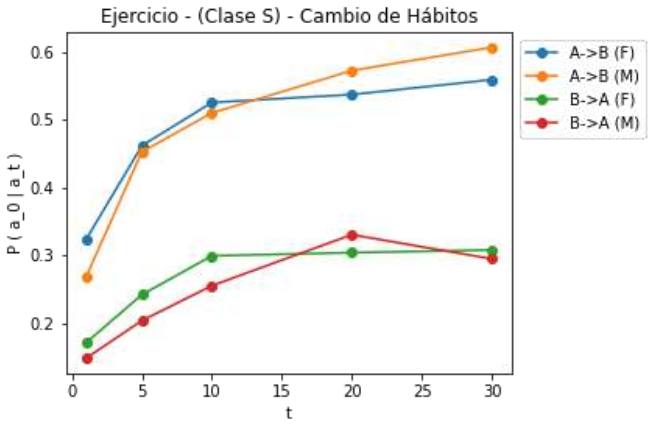


Figura 3.25. Gráfico de las fuerzas de cambio de hábitos de la variable *Ejercicio,* correspondientes a la división de la población *F vs. M.*

Figura 3.26. *Funnel* de cinco etapas de preservación de hábitos buenos de *Ejercicio*, correspondiente a la división de la población *F* vs. *M.*

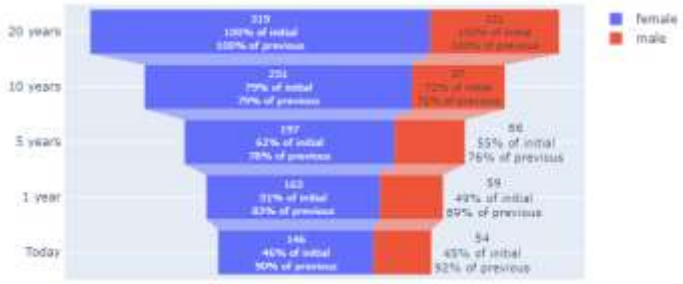


Figura 3.27. *Funnel* de cinco etapas de preservación de hábitos malos de *Ejercicio*, correspondiente a la división de la población *F* vs. *M.*

3.2.2.2 – *Condición Física*

En primer lugar, observamos que, de forma similar a como ocurrió con la variable *Ejercicio*, la fuerza de preservación de hábitos de *Condición Física* disminuye cuando aumenta el tiempo de referencia, es decir, es más factible preservar hábitos (buenos o malos) en el corto plazo (< 5 años) que en el mediano o largo plazo (> 5 años). Asimismo, la fuerza de cambio de hábitos aumenta cuando aumenta el tiempo de referencia, es decir, es más factible observar un cambio hábitos de ejercicio en el largo plazo que en el corto. Este comportamiento se ilustra en las figuras 3.28 y 3.29.

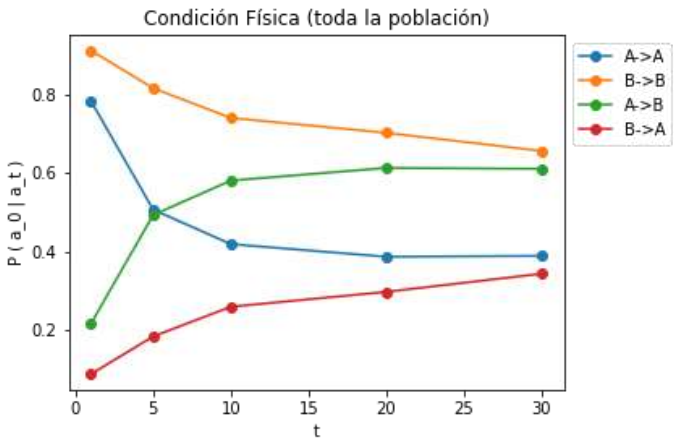


Figura 3.28. Gráfico de las fuerzas de preservación/cambio de hábitos de la variable *Condición Física,* correspondientes a la población general a lo largo de 30 años.

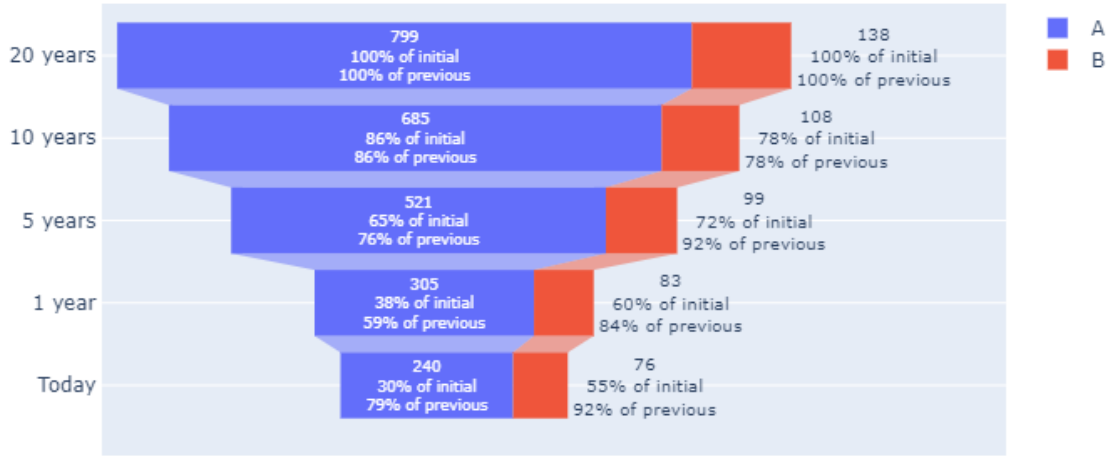


Figura 3.29. Funnel de cinco etapas de preservación de hábitos de *Condición Física* correspondientes a la población general.

De manera similar a lo observado en el análisis correspondiente a la variable *Ejercicio*, la división de clases *O, H* y *A* mostró diferencias importantes entre los comportamientos asociados a diferentes subgrupos de la población. Por un lado, las personas con obesidad (clase *O*) mostraron ser, históricamente, considerablemente más propensas a mantener malos hábitos relacionados a su condición física y menos propensos a mantener buenos hábitos con respecto de las personas sin obesidad (clase *NO*). Del mismo modo, las personas pertenecientes a la clase *O* muestran a ser más propensas a cambiar buenos hábitos por malos, y menos propensas a cambiar malos hábitos por buenos, que las personas de la clase *NO*. Este comportamiento se ilustra en las figuras 3.30 – 3.33.

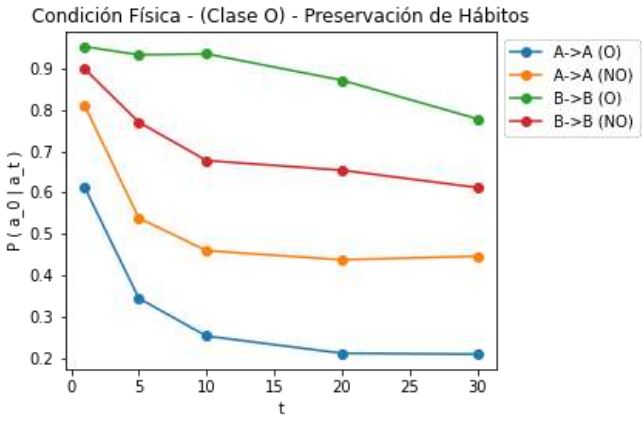


Figura 3.30. Gráfico de las fuerzas de preservación de hábitos de la variable *Condición Física,* correspondientes a la división de la población *O vs. NO*.

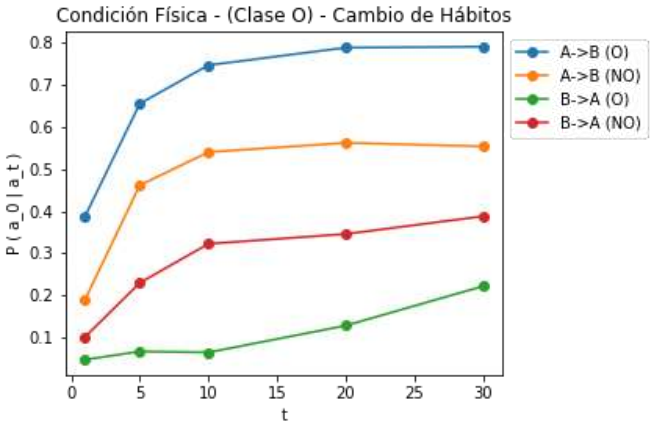


Figura 3.31. Gráfico de las fuerzas de cambio de hábitos de la variable *Condición Física,* correspondientes a la división de la población *O vs. NO*.

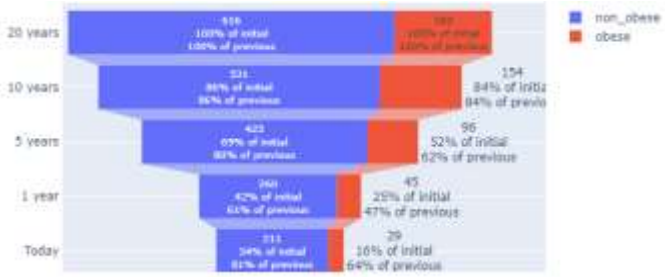


Figura 3.32. *Funnel* de cinco etapas de preservación de hábitos buenos de *Condición Física*, correspondiente a la división de la población *O* vs. *NO*.

Figura 3.33. *Funnel* de cinco etapas de preservación de hábitos malos de *Condición Física*, correspondiente a la división de la población *O* vs. *NO*.

En términos de la clase *H* (figuras 3.34 - 3.37) las personas con alto grado de estudios (clase *H*) mostraron ser históricamente más propensas a mantener buenos hábitos relacionados a *Condición Física* y menos propensas a mantener malos hábitos que las personas sin alto grado de estudios (clase *NH*). Del mismo modo, éstas (clase *H*) muestran ser menos propensas a cambiar un buen hábito por uno malo, y más propensas a cambiar un mal hábito por uno bueno que las personas de la clase *NH*.

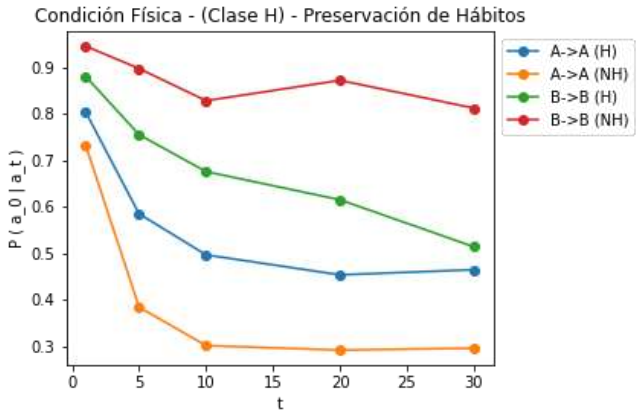


Figura 3.34. Gráfico de las fuerzas de preservación de hábitos de la variable *Condición Física,* correspondientes a la división de la población *H vs. NH.*

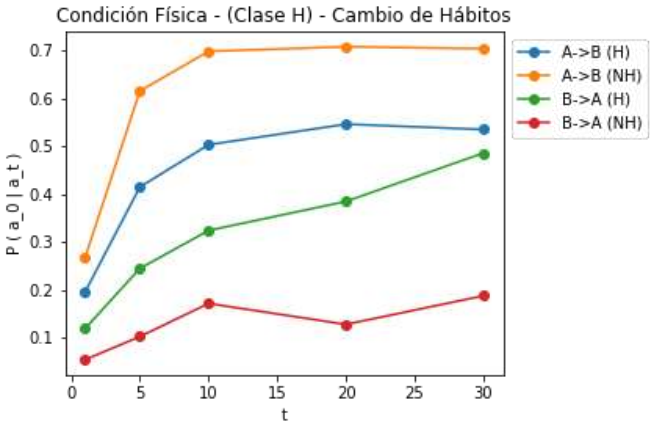


Figura 3.35. Gráfico de las fuerzas de cambio de hábitos de la variable *Condición Física,* correspondientes a la división de la población *H vs. NH.*

Figura 3.36. *Funnel* de cinco etapas de preservación de hábitos buenos de *Condición Física*, correspondiente a la división de la población *H* vs. *NH.*

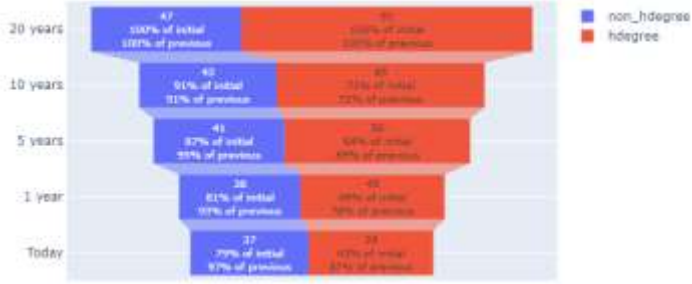


Figura 3.37. *Funnel* de cinco etapas de preservación de hábitos malos de *Condición Física*, correspondiente a la división de la población *H* vs. *NH.*

En el contexto de la división dada por la clase A (figuras 3.38 – 3.41), nuevamente, de manera similar a como ocurrió con la variable *Ejercicio*, las personas con puesto de académico (clase A) mostraron ser históricamente más propensas a mantener buenos hábitos, pero mostraron, al mismo tiempo tener patrones similares de manutención de malos hábitos con respecto a las personas de la clase NA a lo largo de todos los tiempos de referencia (0 – 30 años). Es decir, se refuerza la noción de que la única diferencia de comportamiento de los académicos contra la población no académica parece ser que los académicos abandonan un hábito que ya es bueno con menor frecuencia que los no académicos. Nuevamente, este fenómeno sugiere que la gente académica no necesariamente persigue buenos hábitos sólo porque éstos sean "buenos", sino que simplemente son más constantes en la preservación de los hábitos que ya tienen.

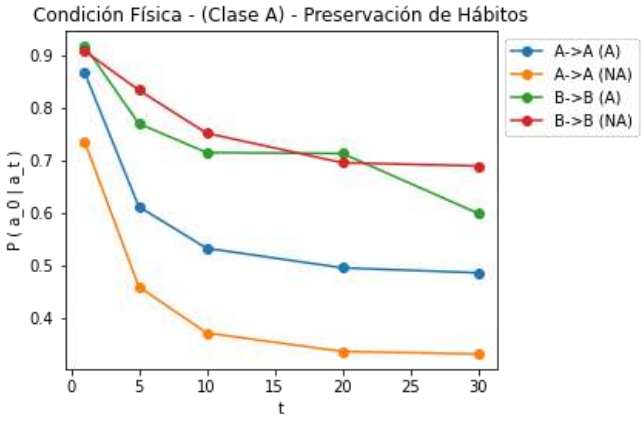


Figura 3.38. Gráfico de las fuerzas de preservación de hábitos de la variable *Condición Física,* correspondientes a la división de la población *A vs. NA*.

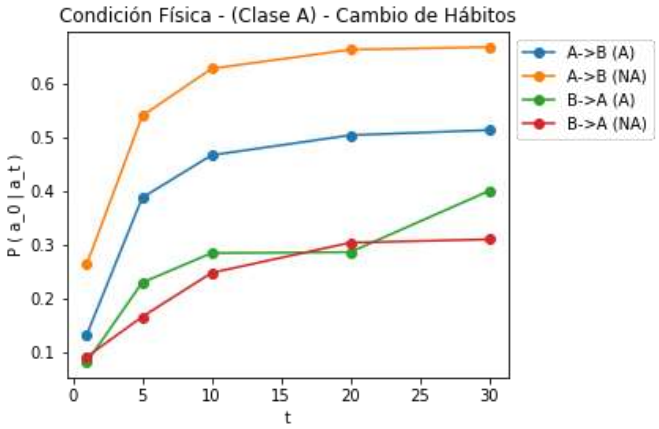


Figura 3.39. Gráfico de las fuerzas de cambio de hábitos de la variable *Condición Física,* correspondientes a la división de la población *A vs. NA*.

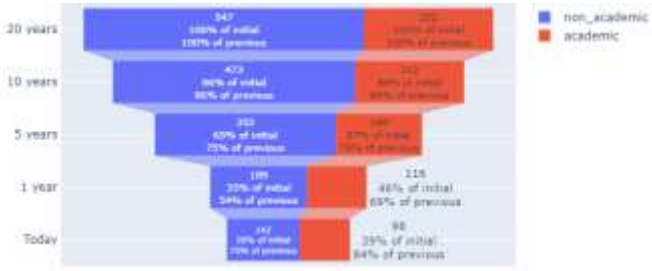


Figura 3.40. *Funnel* de cinco etapas de preservación de hábitos buenos de *Condición Física*, correspondiente a la división de la población *A* vs. *NA*.

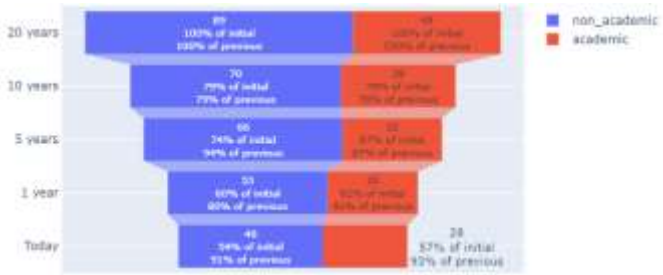


Figura 3.41. *Funnel* de cinco etapas de preservación de hábitos malos de *Condición Física*, correspondiente a la división de la población *A* vs. *NA*.

Como se observa, en general puede decirse que a través de las divisiones de clases *O, H* y *A*, los historiales de *Condición Física* parecen, en sus tendencias y magnitudes, muy similares a los historiales de *Ejercicio,* lo que supone una estrecha relación entre ambos historiales. Intuitivamente esto tiene sentido, pues, como se mencionó anteriormente, la condición física depende de los hábitos de ejercicio y alimentación de las personas. No obstante, al comparar los historiales de *Condición Física* y *Ejercicio* entre hombres y mujeres (figuras 3.42 – 3.45), hay una diferencia notable: en términos de la variable *Condición Física*, los hombres muestran una mayor preservación de buenos hábitos y un menor abandono de buenos hábitos, mientras que, en términos de *Ejercicio* (figuras 3.24 – 3.27), la evolución de los hábitos a través del tiempo es casi indistinta entre los dos. Esto supone la presencia de un factor distintivo entre hombres y mujeres, que podría ser explicado a través de distintas hipótesis:

- Una diferencia hormonal (i.e. una mayor presencia de testosterona hace que los hombres puedan obtener mejores beneficios del ejercicio que las mujeres).

- Una diferencia conductual (i.e. los hombres mantienen mejores dietas que las mujeres).
- Una diferencia de auto percepción (i.e. los hombres sobreestiman su estado de condición física en mayor medida que las mujeres).

3.2.2.3 – Salud

Lo primero que debe notarse al respecto del análisis de los historiales de *Salud* es que en general, las personas parecen reportar, independientemente del grupo al que pertenezcan, que su salud es buena. Es decir, hay un desbalance de respuestas muy sesgado hacia reportes del tipo "A" de la variable *Salud*. Este sesgo se aprecia particularmente en la figura 3.47. Además, como se observa en la figura 3.46, los patrones de conservación de hábitos buenos y los patrones de conversación de hábitos malos se ven prácticamente iguales a lo largo del tiempo de referencia completo (0 – 30 años). Estas dos condiciones nos hablan de que las condiciones de la variable *Salud* de las personas son muy estables cuando se les comparan contra otras variables de *Autoevaluación de Salud*. Dicho de otro modo, la población en general es igualmente propensa a mantener hábitos buenos de salud como a mantener hábitos malos (de la misma manera, son igualmente propensos a cambiar de hábitos buenos a malos, que de malos a buenos).

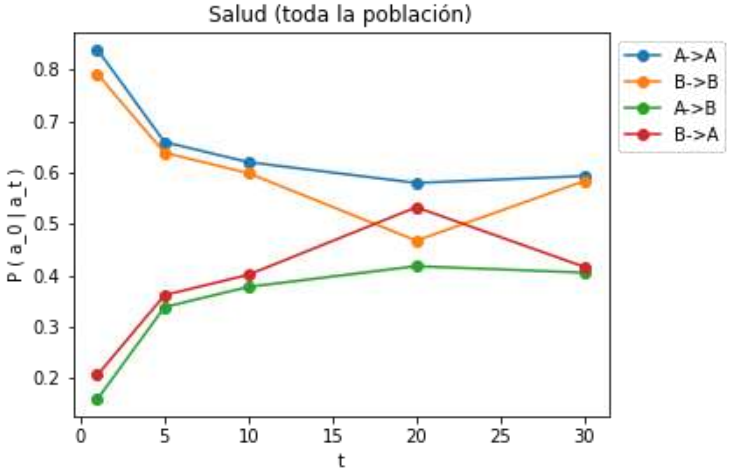


Figura 3.46. Gráfico de las fuerzas de preservación/cambio de hábitos de la variable *Salud,* correspondientes a la población general a lo largo de 30 años.

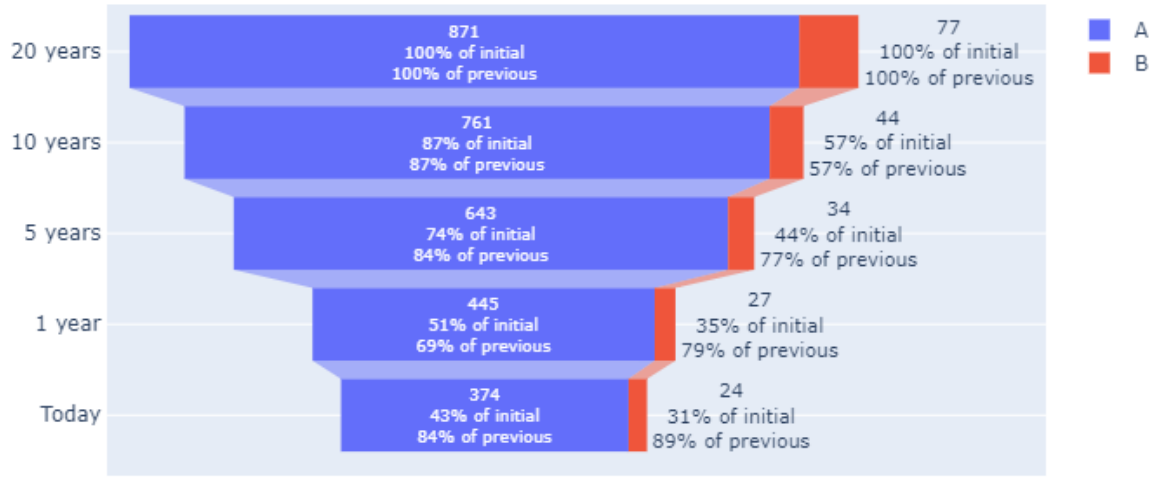


Figura 3.47. Funnel de cinco etapas de preservación de hábitos de *Salud* correspondientes a la población general.

Asimismo, a lo largo de todas las clases analizadas, observamos que, de forma similar a como ocurrió con la variable *Ejercicio* y *Condición Física*, la fuerza de preservación de hábitos disminuye cuando aumenta el tiempo de referencia, es decir, es más factible preservar hábitos relacionados a *Salud* (buenos o malos) en el corto plazo (< 5 años) que en el mediano o largo plazo (> 5 años). Asimismo, la fuerza de cambio de hábitos aumenta cuando aumenta el tiempo de referencia, es decir, es más factible cambiar hábitos de ejercicio en el largo plazo que en el corto.

En el contexto específico de la división de personas según su obesidad, las personas con obesidad (clase *O*) mostraron ser, históricamente, considerablemente más propensas a mantener malos hábitos relacionados a su salud y menos propensos a mantener buenos hábitos con respecto de las personas sin obesidad (clase *NO*). Del mismo modo, las personas pertenecientes a la clase *O* muestran a ser más propensas a cambiar buenos hábitos por malos, y menos propensas a cambiar malos hábitos por buenos, que las personas de la clase *NO*. Este comportamiento se ve ilustrado en las figuras 3.48 – 3.49.

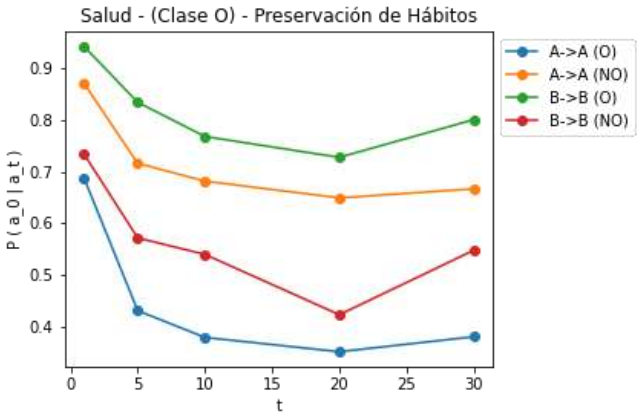


Figura 3.48. Gráfico de las fuerzas de preservación de hábitos de la variable *Salud,* correspondientes a la división de la población *O vs. NO*.

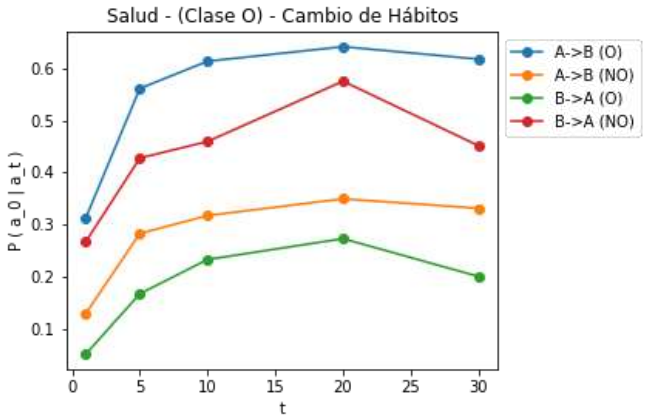


Figura 3.49. Gráfico de las fuerzas de cambio de hábitos de la variable *Salud,* correspondientes a la división de la población *O vs. NO*.

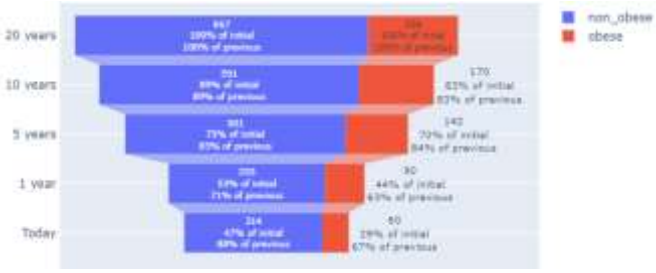


Figura 3.50. *Funnel* de cinco etapas de preservación de hábitos buenos de *Salud*, correspondiente a la división de la población *O* vs. *NO*.

Figura 3.51. *Funnel* de cinco etapas de preservación de hábitos malos de *Salud*, correspondiente a la división de la población *O* vs. *NO*.

En el caso de la división de la población dada por H/NH, no se observó ninguna diferencia aparente respecto a los hábitos mostrados a lo largo del tiempo (figuras 3.52 – 3.55). No obstante, la división de la población dada por A/NA sí marcó una diferencia importante de comportamiento: las personas con puesto académico (clase *A*) mostraron ser, históricamente, considerablemente menos propensas a mantener malos hábitos relacionados a su salud y más propensos a mantener buenos hábitos con respecto de las personas no académicas (clase *NA*). Del mismo modo, las personas pertenecientes a la clase *A* muestran a ser más propensas a cambiar buenos hábitos por malos, y menos propensas a cambiar malos hábitos por buenos, que las personas de la clase *NA* (figuras 3.56 – 3.59).

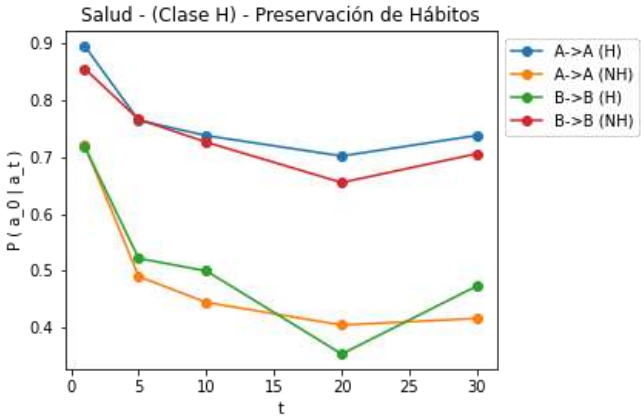


Figura 3.52. Gráfico de las fuerzas de preservación de hábitos de la variable *Salud,* correspondientes a la división de la población *H vs. NH*.

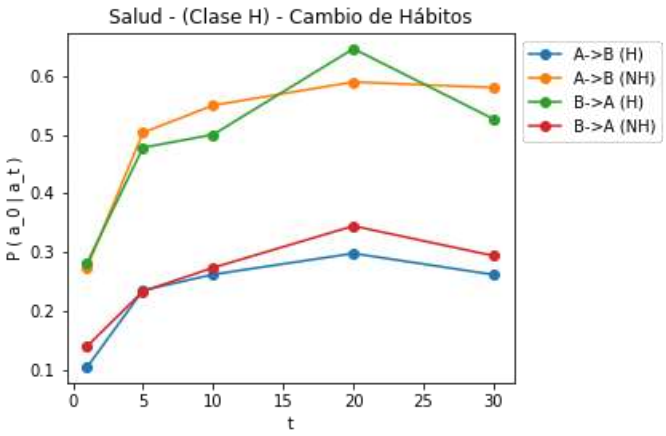


Figura 3.53. Gráfico de las fuerzas de cambio de hábitos de la variable *Salud,* correspondientes a la división de la población *H vs. NH*.

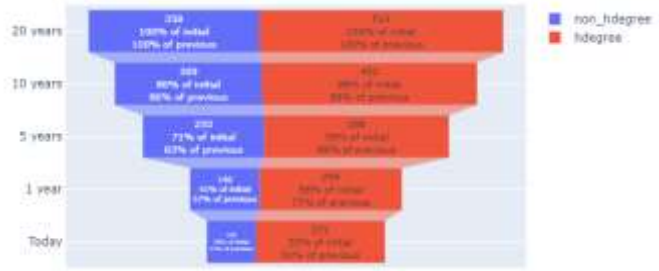


Figura 3.54. *Funnel* de cinco etapas de preservación de hábitos buenos de *Salud*, correspondiente a la división de la población *H* vs. *NH*.

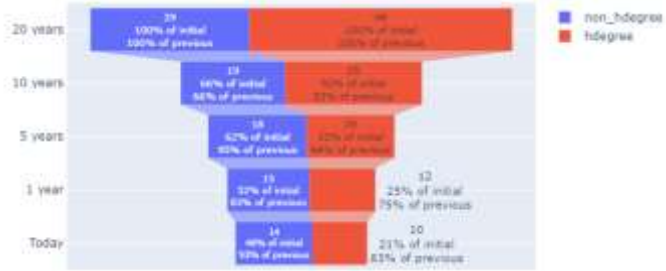


Figura 3.55. *Funnel* de cinco etapas de preservación de hábitos malos de *Salud*, correspondiente a la división de la población *H* vs. *NH*.

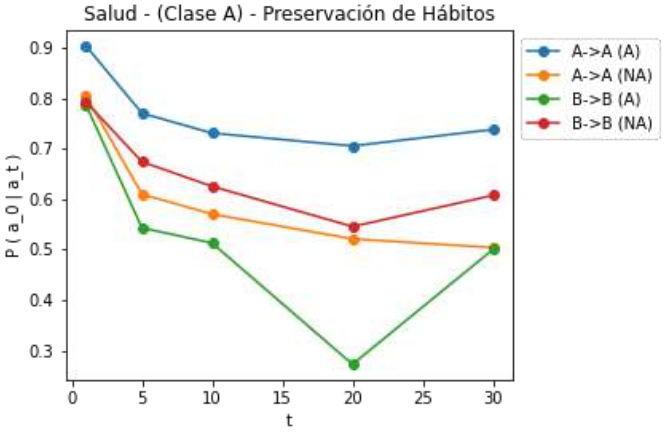


Figura 3.56. Gráfico de las fuerzas de preservación de hábitos de la variable *Salud,* correspondientes a la división de la población *A vs. NA*.

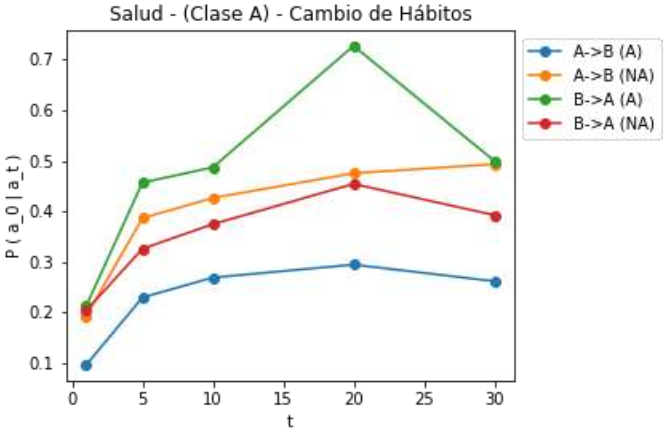


Figura 3.57. Gráfico de las fuerzas de cambio de hábitos de la variable *Salud,* correspondientes a la división de la población *A vs. NA*.

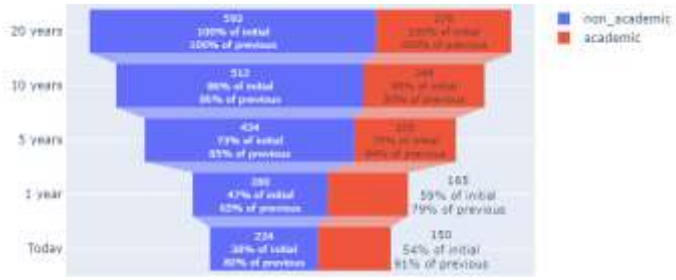


Figura 3.58. *Funnel* de cinco etapas de preservación de hábitos buenos de *Salud*, correspondiente a la división de la población *A* vs. *NA*.

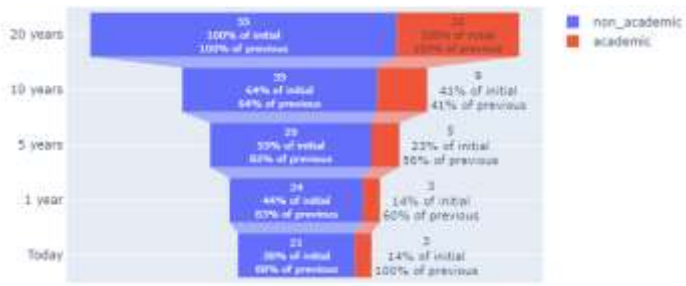


Figura 3.59. *Funnel* de cinco etapas de preservación de hábitos malos de *Salud*, correspondiente a la división de la población *A* vs. *NA*.

Por último, los hombres mostraron, históricamente, ser más propensos a mantener buenos hábitos relacionados a *Salud* que las mujeres, y menos propensos a mantener malos hábitos. La única excepción es el plazo de 1 año, donde los patrones de preservación/cambio son prácticamente los mismos para hombres y mujeres (figuras 3.60 – 3.63).

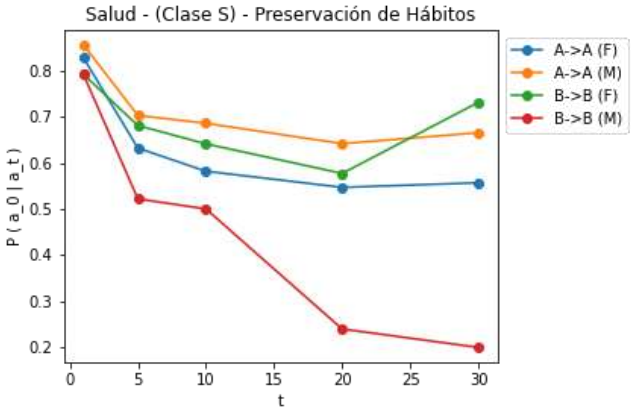


Figura 3.60. Gráfico de las fuerzas de preservación de hábitos de la variable *Salud,* correspondientes a la división de la población *F vs. M.*

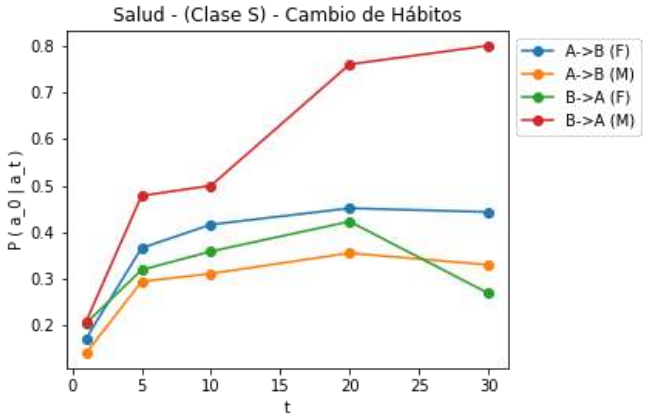


Figura 3.61. Gráfico de las fuerzas de cambio de hábitos de la variable *Salud,* correspondientes a la división de la población *F vs. M.*

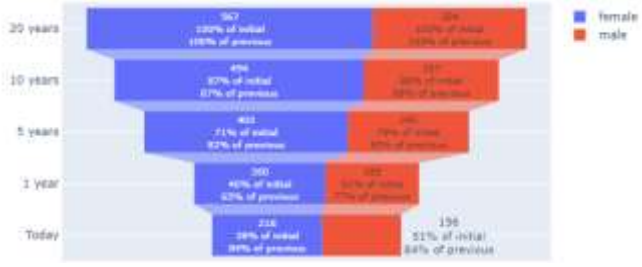


Figura 3.62. *Funnel* de cinco etapas de preservación de hábitos buenos de *Salud*, correspondiente a la división de la población *F* vs. *M*.

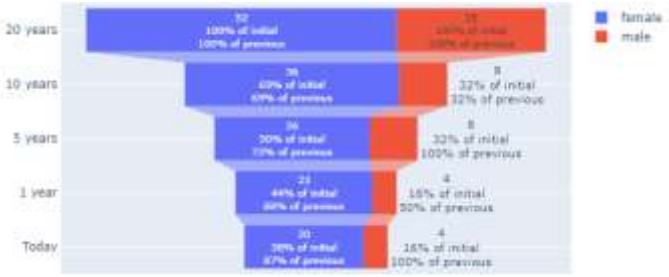


Figura 3.63. *Funnel* de cinco etapas de preservación de hábitos malos de *Salud*, correspondiente a la división de la población *F* vs. *M*.

3.2.2.4 – Estrés

Contrariamente a los reportes de *Salud*, la población en general parece reportar que su condición de *Estrés* es mala, es decir, hay un desbalance de respuestas muy sesgado hacia reportes del tipo "B" (figura 3.65). Además, los patrones de conservación de hábitos buenos y los patrones de conversación de hábitos malos siguen tendencias muy similares (figura 3.64) a lo largo del tiempo de referencia completo (0 – 30 años). Estas dos condiciones nos hablan de que las condiciones de la variable *Estrés* de las personas son muy estables cuando se les comparan contra otras variables como *Ejercicio* o *Condición Física*.

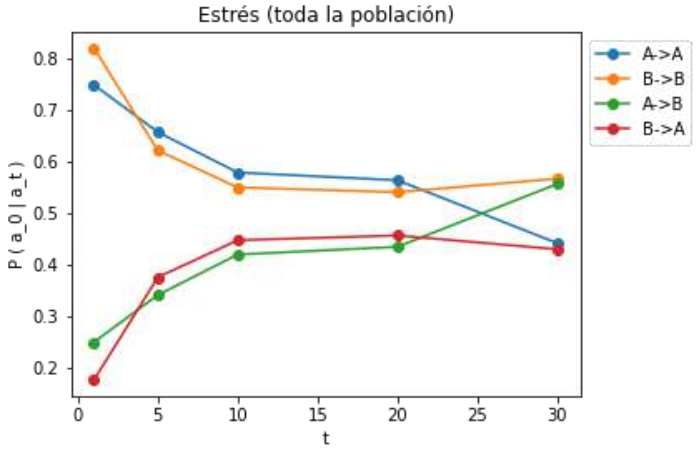


Figura 3.64. Gráfico de las fuerzas de preservación/cambio de hábitos de la variable *Estrés,* correspondientes a la población general a lo largo de 30 años.

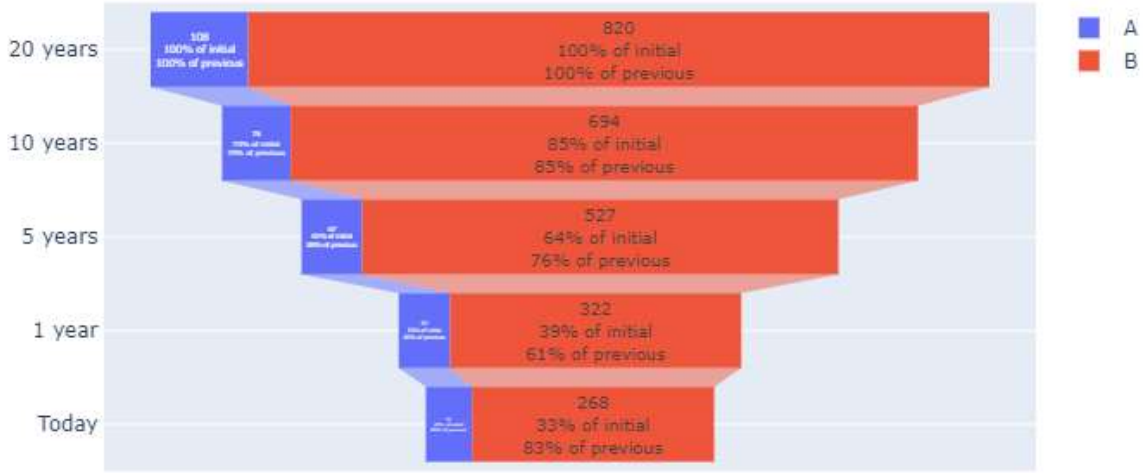


Figura 3.65. Funnel de cinco etapas de preservación de hábitos de *Estrés* correspondientes a la población general.

La división de la clase *O* no marcó ninguna diferencia aparente con respecto a las condiciones de estrés reportadas a lo largo del tiempo (figuras 3.66 – 3.69). Esto puede ser una evidencia de que el estrés está más relacionado con condiciones externas (trabajo,

dinero, etcétera) [13] que con hábitos de vida de una persona pues, hasta ahora, todas las demás variables de *Autoevaluación de Salud* habían causado una división muy marcada entre los comportamientos de la clase *O* y la clase *NO*.

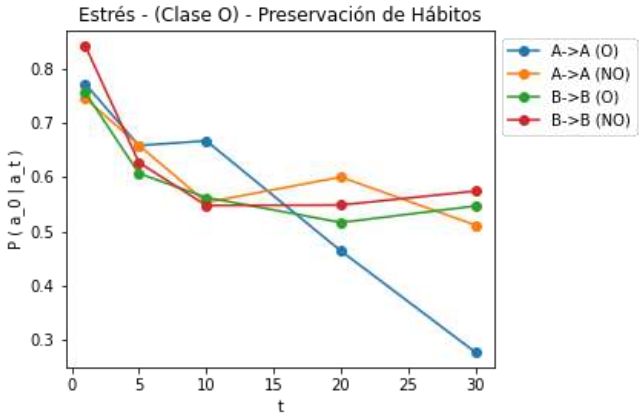


Figura 3.66. Gráfico de las fuerzas de preservación de hábitos de la variable *Estrés,* correspondientes a la división de la población *O vs. NO*.

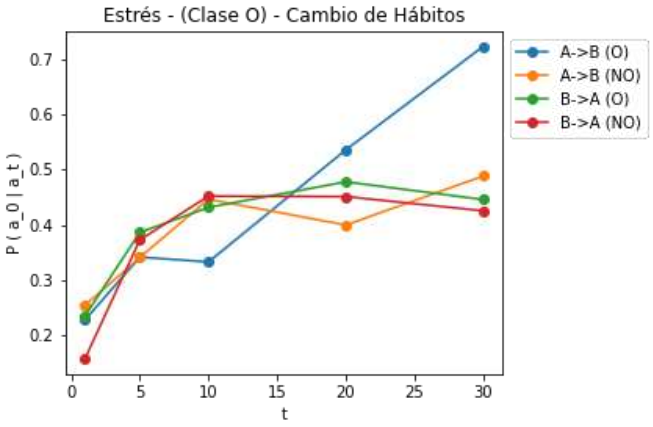


Figura 3.67. Gráfico de las fuerzas de cambio de hábitos de la variable *Estrés,* correspondientes a la división de la población *O vs. NO*.

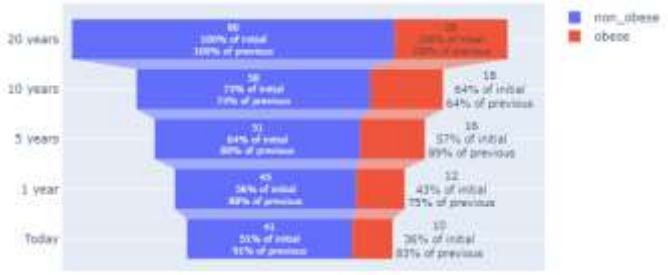


Figura 3.68. *Funnel* de cinco etapas de preservación de hábitos buenos de *Estrés*, correspondiente a la división de la población *O* vs. *NO*.

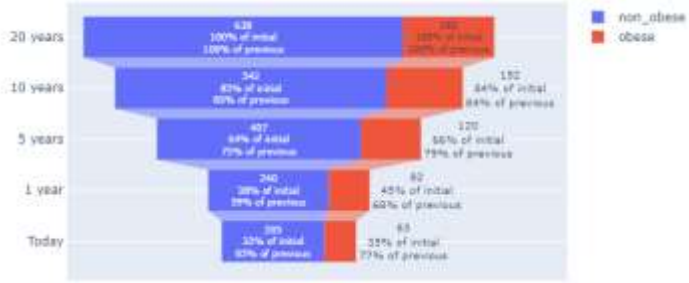


Figura 3.69. *Funnel* de cinco etapas de preservación de hábitos malos de *Estrés*, correspondiente a la división de la población *O* vs. *NO*.

Por otra parte, la división de la clase *H* no marcó ninguna diferencia aparente con respecto a los hábitos mostrados a lo largo del tiempo (figuras 3.70 – 3.73). No obstante, con respecto a la división de la clase *A*, las personas con puesto académico (clase *A*) mostraron ser históricamente más propensas a mantener una buena condición de *Estrés* pero, interesantemente, más propensos también a mantener una mala condición de *Estrés* con respecto de las personas no académicas (clase *NA*). Este resultado, sin embargo, parece más bien causa de que la evidencia asociada a patrones malos ("B") de *Estrés* es muy poca para la intersección entre las personas que reportan una condición de estrés mala y las personas académicas.

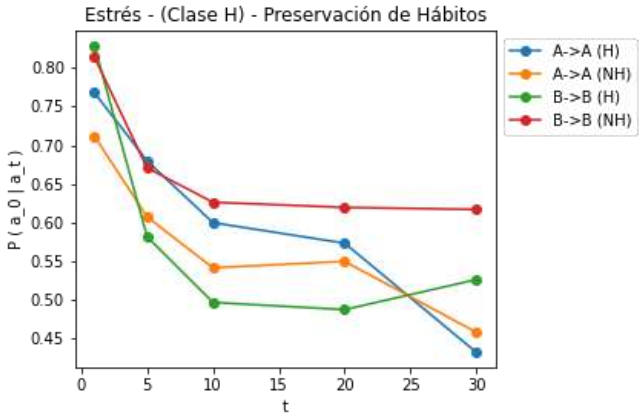


Figura 3.70. Gráfico de las fuerzas de preservación de hábitos de la variable *Estrés,* correspondientes a la división de la población *H vs. NH.*

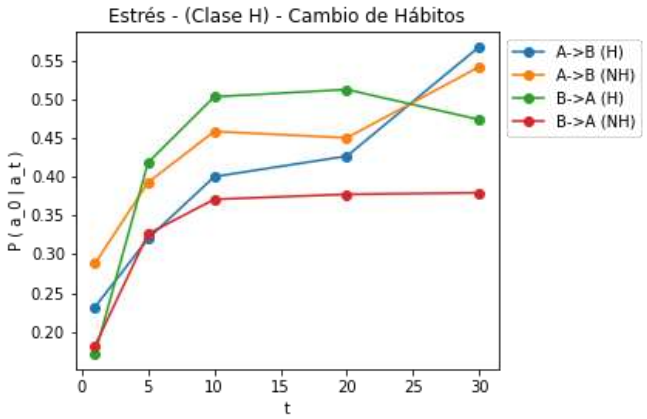


Figura 3.71. Gráfico de las fuerzas de cambio de hábitos de la variable *Estrés,* correspondientes a la división de la población *H vs. NH.*

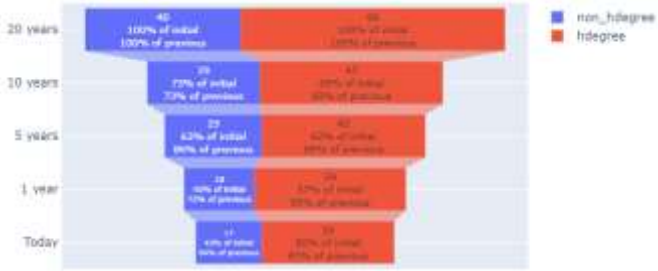


Figura 3.72. *Funnel* de cinco etapas de preservación de hábitos buenos de *Estrés*, correspondiente a la división de la población *H* vs. *NH.*

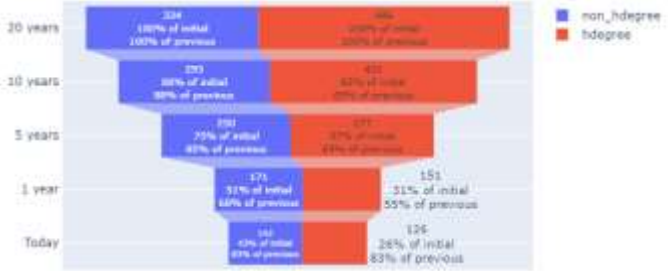


Figura 3.73. *Funnel* de cinco etapas de preservación de hábitos malos de *Estrés*, correspondiente a la división de la población *H* vs. *NH.*

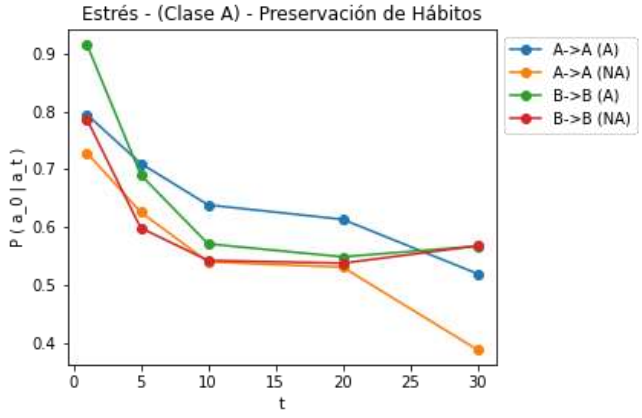


Figura 3.74. Gráfico de las fuerzas de preservación de hábitos de la variable *Estrés,* correspondientes a la división de la población *A vs. NA.*

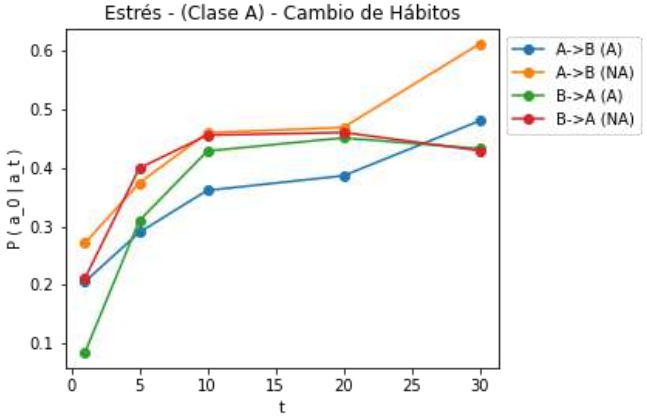


Figura 3.75. Gráfico de las fuerzas de cambio de hábitos de la variable *Estrés,* correspondientes a la división de la población *A vs. NA.*

Figura 3.76. *Funnel* de cinco etapas de preservación de hábitos buenos de *Estrés*, correspondiente a la división de la población *A* vs. *NA*.

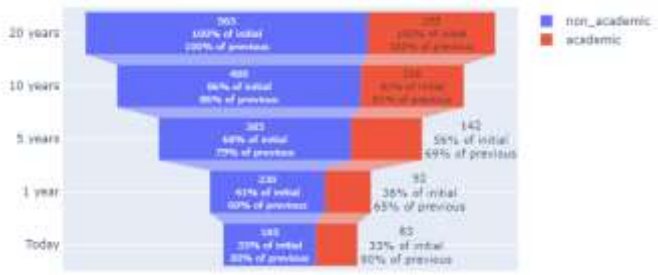


Figura 3.77. *Funnel* de cinco etapas de preservación de hábitos malos de *Estrés*, correspondiente a la división de la población *A* vs. *NA*.

3.2.2.5 – Peso

En la población en general, se observa que los patrones de las fuerzas de preservación y cambio relacionados con el historial de la variable *Peso* son marcadamente diferentes. Lo más notable es que para esta variable, la fuerza de preservación de una mala condición es prácticamente igual a 1 (es decir, las personas que tienen una mala condición de peso están fuertemente atadas a mantener dicha condición a lo largo del tiempo). La fuerza de preservación de un buen hábito disminuye conforme aumenta el tiempo de referencia, pero ésta se vuelve considerablemente pequeña (con relación a otras variables analizadas) a partir de los 5 años. Esto quiere decir que es muy difícil para las personas mantener un buen peso, incluso cuando el tiempo de referencia es muy pequeño, y es muy fácil mantener una mala condición de peso, incluso cuando el tiempo de referencia es muy grande. Esto se ilustra en las figuras 3.82 y 3.83.

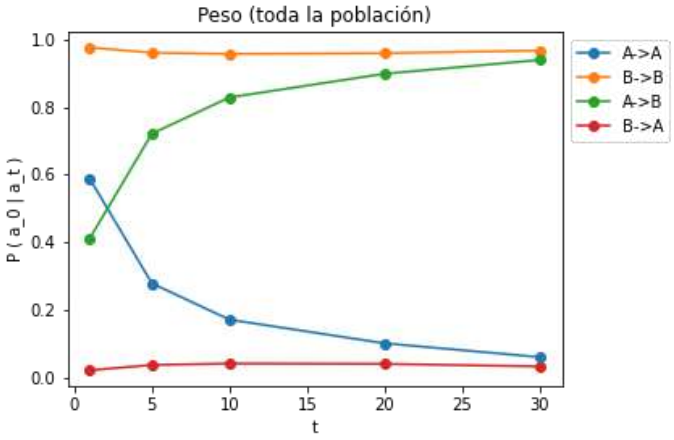


Figura 3.82. Gráfico de las fuerzas de preservación/cambio de hábitos de la variable *Peso,* correspondientes a la población general a lo largo de 30 años.

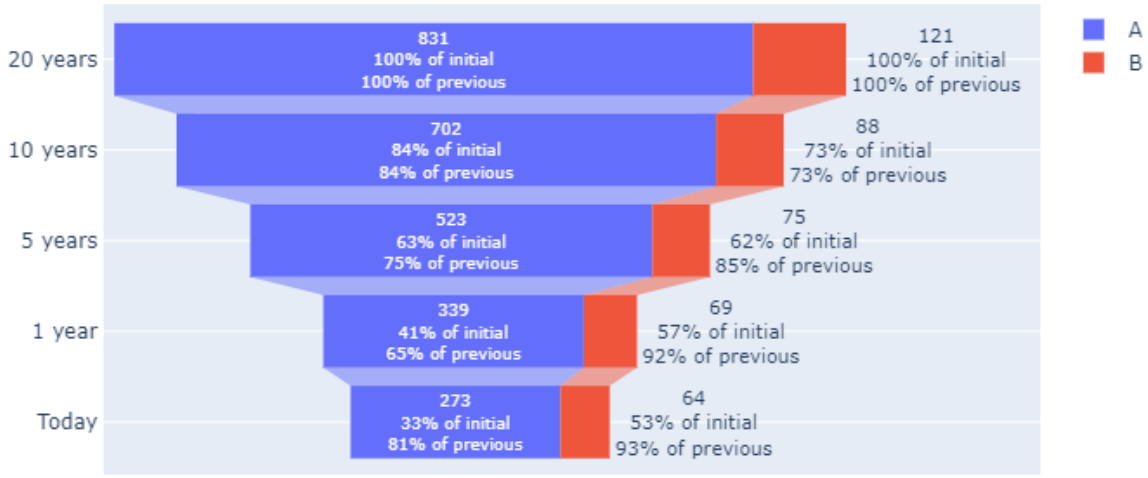


Figura 3.83. Funnel de cinco etapas de preservación de hábitos de *Peso* correspondientes a la población general.

Al analizar la división de clase *O*, se observa que, si bien hay una clara diferencia entre la preservación de una condición buena de peso entre las personas con obesidad y sin obesidad, no hay mucha diferencia a la hora de hablar de preservación de una mala condición de peso a lo largo del tiempo (figuras 3.84 – 3.87). Es decir, sin importar si hoy una persona tiene o no obesidad, si la persona reportó un peso malo en el pasado, será mucho más propensa a reportar un peso malo actualmente. Para las personas que no tienen obesidad hoy en día, y que están reportando un peso malo a lo largo del tiempo, este fenómeno puede explicarse a partir del hecho de que, en general, las personas tienden a considerar su peso como peor de lo que en realidad es.

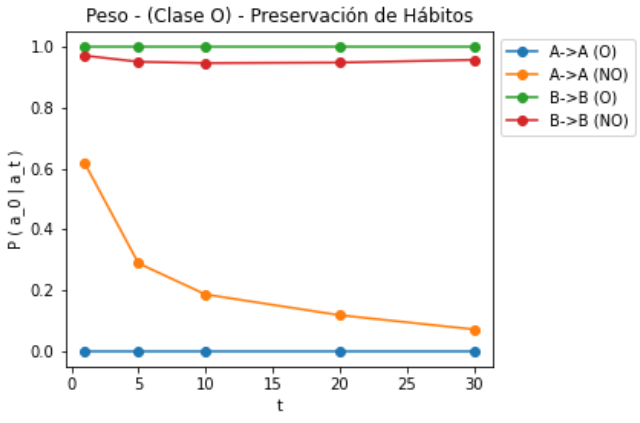


Figura 3.84. Gráfico de las fuerzas de preservación de hábitos de la variable *Peso,* correspondientes a la división de la población *O vs. NO.*

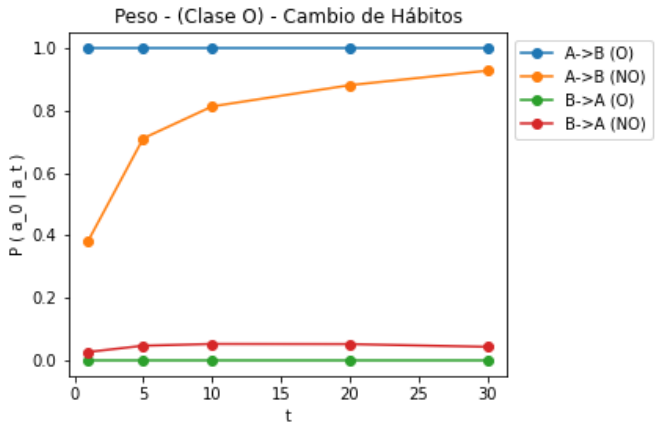


Figura 3.85. Gráfico de las fuerzas de cambio de hábitos de la variable *Peso,* correspondientes a la división de la población *O vs. NO.*

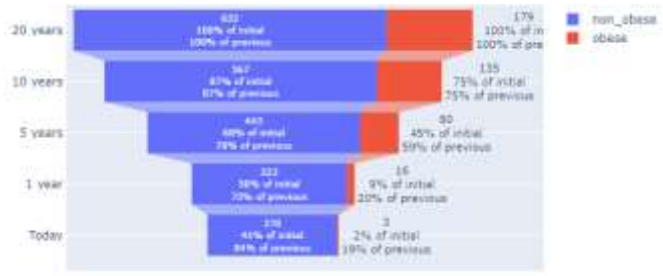


Figura 3.86. *Funnel* de cinco etapas de preservación de hábitos buenos de *Peso*, correspondiente a la división de la población *O* vs. *NO*.

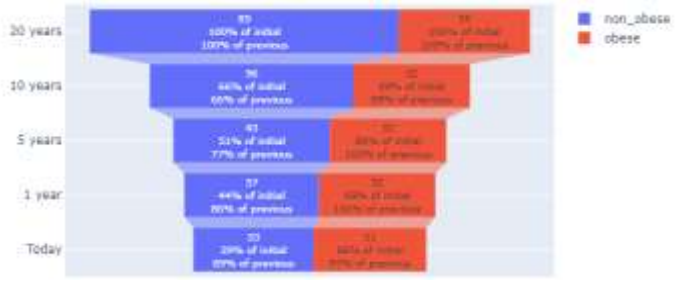


Figura 3.87. *Funnel* de cinco etapas de preservación de hábitos malos de *Peso*, correspondiente a la división de la población *O* vs. *NO*.

La división de la clase *H* no marcó ninguna diferencia aparente con respecto a las condiciones de peso reportadas a lo largo del tiempo (figuras 3.88 – 3.91). Por otro lado, hablando de la división de clase *A*, las personas con puesto académico (clase *A*) mostraron ser, históricamente, menos propensas a mantener una buena condición de peso, pero mostraron ser igualmente propensas a mantener una mala condición de peso con respecto de las personas de la población *NA* (figuras 3.92 – 3.95). Esta es una observación interesante respecto a la clase A, a la que habíamos asociado en secciones anteriores a una mejor preservación de buenos hábitos de *Ejercicio, Salud* y *Condición Física* a lo largo del tiempo con respecto de la población *NA*. Esto podría explicarse como que las personas de la clase *A* son más severas a la hora de juzgar su propio peso que la clase *NA*, pero que no necesariamente esto implica que esa evaluación es cierta.

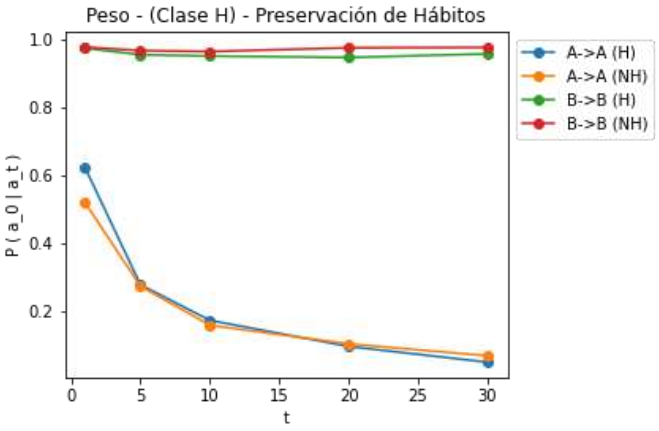


Figura 3.88. Gráfico de las fuerzas de preservación de hábitos de la variable *Peso,* correspondientes a la división de la población *H vs. NH*.

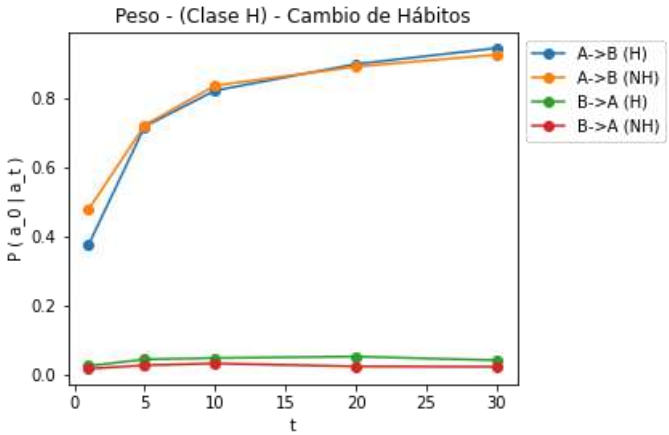


Figura 3.89. Gráfico de las fuerzas de cambio de hábitos de la variable *Peso,* correspondientes a la división de la población *H vs. NH*.

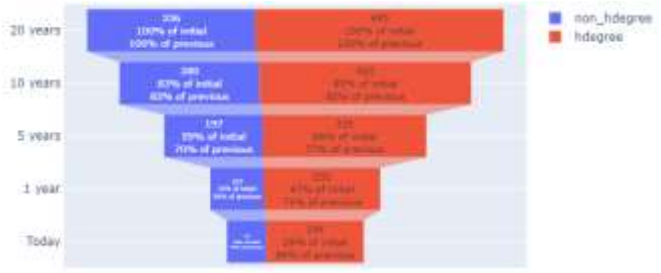


Figura 3.90. *Funnel* de cinco etapas de preservación de hábitos buenos de *Peso*, correspondiente a la división de la población *H* vs. *NH*.

Figura 3.91. *Funnel* de cinco etapas de preservación de hábitos malos de *Peso*, correspondiente a la división de la población *H* vs. *NH*.

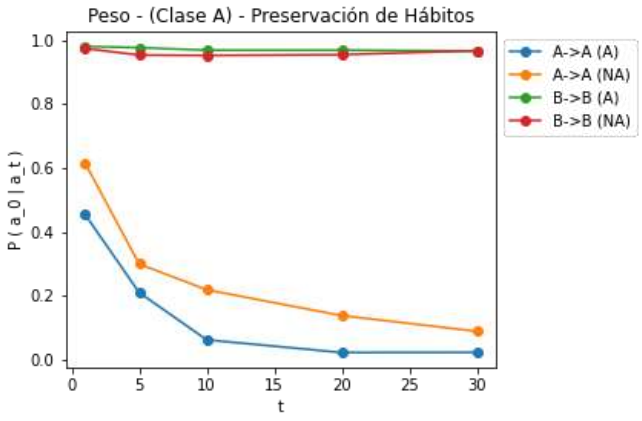


Figura 3.92. Gráfico de las fuerzas de preservación de hábitos de la variable *Peso,* correspondientes a la división de la población *A vs. NA*.

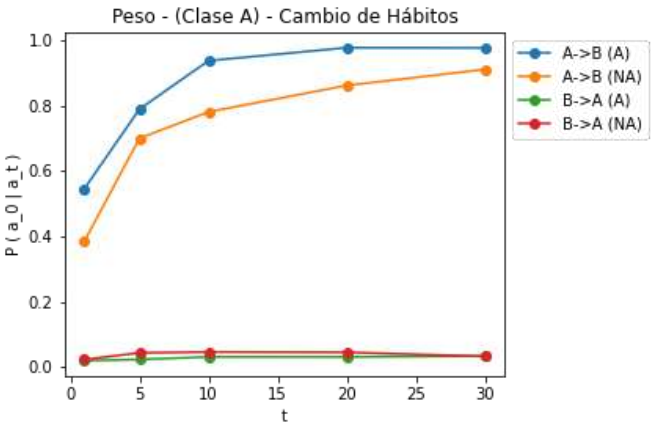


Figura 3.93. Gráfico de las fuerzas de cambio de hábitos de la variable *Peso,* correspondientes a la división de la población *A vs. NA*.

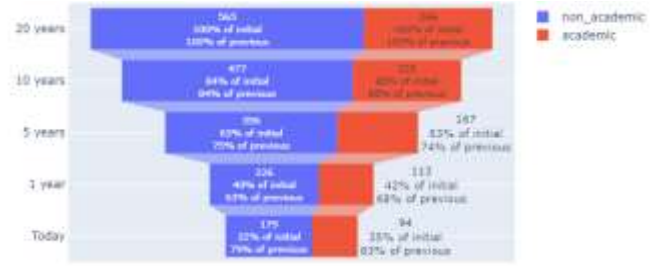


Figura 3.94. *Funnel* de cinco etapas de preservación de hábitos buenos de *Peso*, correspondiente a la división de la población *A* vs. *NA*.

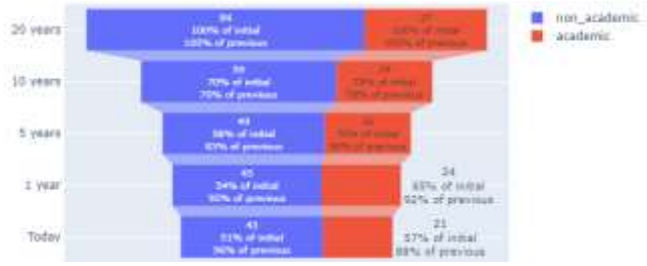


Figura 3.95. *Funnel* de cinco etapas de preservación de hábitos malos de *Peso*, correspondiente a la división de la población *A* vs. *NA*.

Siguiendo la tendencia de las otras subpoblaciones mostradas, la división de sexo no marcó diferencia en términos de la preservación de una mala condición de peso: tanto hombres como mujeres mostraron ser igualmente propensos a mantener una mala condición de peso a lo largo del tiempo. Es decir, es igualmente difícil para la población general, independientemente de los subgrupos a los que pertenezca, cambiar una condición mala

de peso por una buena. No obstante, los hombres parecen ser más propensos a mantener una buena condición de peso que las mujeres a lo largo del tiempo.

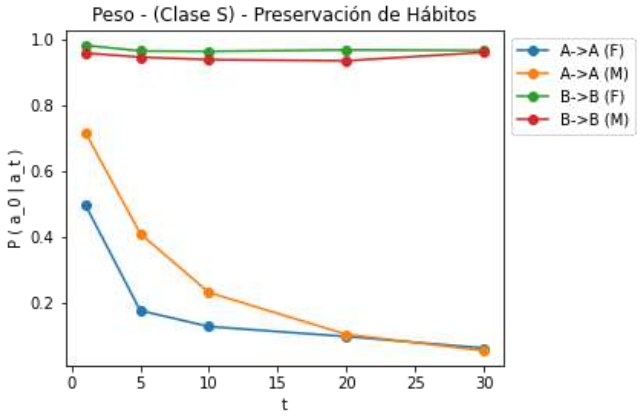


Figura 3.96. Gráfico de las fuerzas de preservación de hábitos de la variable *Peso,* correspondientes a la división de la población *F vs. M.*

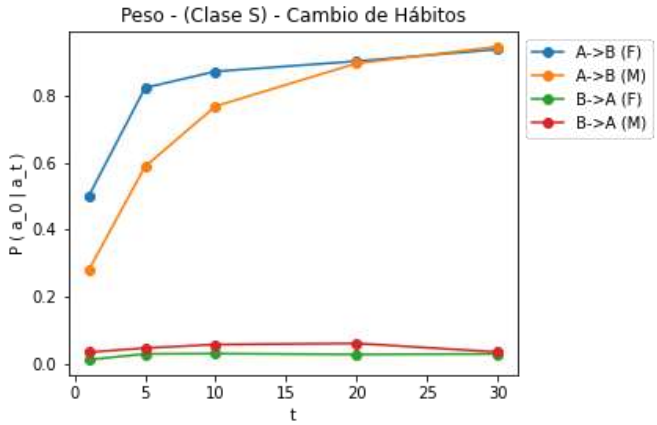


Figura 3.97. Gráfico de las fuerzas de cambio de hábitos de la variable *Peso,* correspondientes a la división de la población *F vs. M.*

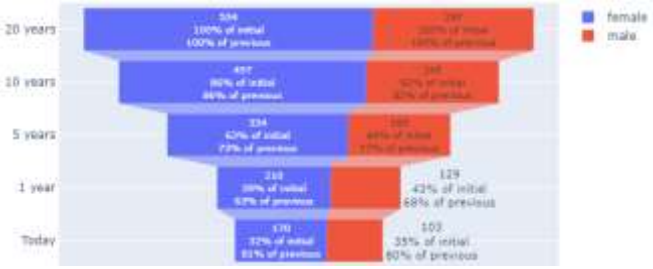


Figura 3.98. *Funnel* de cinco etapas de preservación de hábitos buenos de *Peso*, correspondiente a la división de la población *F* vs. *M.*

Figura 3.99. *Funnel* de cinco etapas de preservación de hábitos malos de *Peso*, correspondiente a la división de la población *F* vs. *M.*

Capítulo 4. Relación entre historias y subgrupos de la población

4.1 – Análisis de épsilons asociados a historiales de hábitos

Tal y como se discutió en la sección 2.2, además de analizar cómo evolucionan en el tiempo de los hábitos relacionados con las variables de *Autoevaluación de Salud* en la población general y en diferentes subgrupos de la población, queremos también perfilar qué tipo de historiales de hábitos están asociados a estos mismos subgrupos de la población. Es decir, queremos perfilar a estos subgrupos de la población, en términos de los tipos de historiales de hábitos que los definen. Para hacer esto, haremos un "Análisis de épsilons" asociados a historiales de hábitos, cuya metodología se describe a continuación.

4.1.1 – Metodología del "Análisis de Épsilons"

Sea G un subgrupo de la población total, dado por las clases *O, H, A* y *S* definidas anteriormente, y sea $\boldsymbol{X}$ un patrón de historial de hábitos asociado a alguna variable de *Autoevaluación de Salud* a, la intención de este análisis es la de encontrar asociaciones entre diferentes G y diferentes $\boldsymbol{X}$. Para encontrar estas asociaciones, necesitamos dos mecanismos:

I. Un mecanismo de representación de $\boldsymbol{X}$. Una forma de generalizar y expresar a cualquier patrón de historias $\boldsymbol{X}$ en términos de la variable a a la que está asociado.
II. Un mecanismo de diagnóstico estadístico. Una medida para establecer qué tan estadísticamente significativa es la relación entre el hecho de que una persona pertenezca a un subgrupo G de la población, y el hecho de que esa persona además muestra un patrón de historial de hábitos $\boldsymbol{X}$ para la variable a en cuestión.

Ambos mecanismos los planteamos a continuación.

4.1.1.1 - Mecanismo de representación de $\boldsymbol{X}$

En términos del punto I., para representar a los patrones de historias $\boldsymbol{X}$, usaremos una representación similar[1] a la utilizada por Stephens et. al en [5]. Sea pues a_t la etiqueta de una variable de *Autoevaluación de Salud* clusterizada según lo descrito en la sección 3.1.1., para un tiempo de referencia t, $\boldsymbol{X}$ será la concatenación de a_t para todos los

[1] La única variante de la representación aquí empleada vs. la propuesta en el artículo citado es el orden de las variables. En aquel artículo, la representación utilizada concatena a las variables en el orden $a_{30}a_{20}a_{10}a_5a_1a_{act}$, donde el subíndice de la variable de *Autoevaluación de Salud* "a" es el tiempo de referencia.

tiempos de referencia disponibles en la base de datos, o sea $t = act, 1, 5, 10, 20\ y\ 30$. Es decir

$$\boldsymbol{X} = a_{act}a_1a_5a_{10}a_{20}a_{30}$$

Donde, $a_t \in \{A, B, N\}$. De este modo, podemos representar a un historial de hábitos relacionados a cada variable de *Autoevaluación de Salud* como sigue:

$$\boldsymbol{X}_{Ejercicio} = ejer\_actB \cdot ejer\_1B \cdot ejer\_5B \cdot ejer\_10B \cdot ejer\_20B \cdot ejer\_30B$$

$$\boldsymbol{X}_{Condición\ Física} = condi\_actB \cdot condi\_1B \cdot condi\_5B \cdot condi\_10B \cdot condi\_20B \cdot condi\_30B$$

$$\boldsymbol{X}_{Salud} = salud\_actB \cdot salud\_1B \cdot salud\_5B \cdot salud\_10B \cdot salud\_20B \cdot salud\_30B$$

$$\boldsymbol{X}_{Estrés} = estres\_actB \cdot estres\_1B \cdot estres\_5B \cdot estres\_10B \cdot estres\_20B \cdot estres\_30B$$

$$\boldsymbol{X}_{Peso} = peso\_actB \cdot peso\_1B \cdot peso\_5B \cdot peso\_10B \cdot peso\_20B \cdot peso\_30B$$

Por ejemplo, para representar el historial de *Ejercicio* de una persona $\boldsymbol{p}$ que reportó un valor "A" de *Ejercicio* actualmente, pero reportó valores de *Ejercicio* "B" los últimos 30 años (i.e. $\boldsymbol{p}$ reportó *ejer_ 0B = A*, *ejer_ 1B = B*, *ejer_ 5B = B*, *ejer_ 10B* = B, *ejer_ 20B = B*, *ejer_ 30B = B*), tendríamos:

$$\boldsymbol{X}^{p}_{Ejercicio} = ABBBBB$$

Un problema de esta representación es que potencialmente ofusca la búsqueda de patrones de historias interesantes, pues hay patrones que son poco probables que ocurran [5]. Por ello, y dado que tenemos muchos tipos de patrones que vamos a buscar asociar a 1076 personas, debemos extender nuestra representación de patrones de historia con el fin de incrementar el número de personas que pueden estar asociadas a un mismo patrón (incrementando así el potencial de análisis de la significancia estadística de dicho patrón). Para este fin, introducimos una notación de Computación Evolutiva [5][14] correspondiente a la representación del *caracter wildcard* "*". Así, la presencia de un "*" dentro de un historial $\boldsymbol{X}$ significará que no nos interesa el valor reportado de la variable de salud para el tiempo de referencia donde esté ubicado posicionalmente ese "*" dentro de la secuencia. Por ejemplo, un patrón de *Condición Física* $\boldsymbol{X}_{Condición\ Física} = A * *** B$ representará a las personas que actualmente reportaron un valor de *Condición Física* bueno (i.e. *condi_ actB = "A"*) y que reportaron un valor de *Condición Física* hace 30 años (i.e. *condi_ 30B = B*), independientemente de lo que hayan reportado en los años de referencia 1, 5, 10 y 20.

Para incrementar la cantidad de personas para las que tengamos historiales completos en un rango de tiempo determinado, sólo analizaremos historiales con información de los últimos 10 años. Para este rango de tiempo, prácticamente toda la población reportó valores válidos ("A" o "B") en sus variables de *Autoevaluación de Salud*. En términos prácticos, sólo analizaremos patrones con la forma general $\boldsymbol{X} = X_1X_2X_3X_4$ **, donde $X \in \{A, B, *\}$. En este sentido, $\boldsymbol{X}$ podrá verse como una variable aleatoria discreta que puede tomar $3^4 = 81$ valores distintos.

4.1.1.2 - Mecanismo de diagnóstico estadístico

Para tener una medida de la relación entre patrones de historias y subgrupos de la población, proponemos usar la *herramienta de diagnóstico épsilon* propuesta por Stephens et. al en [15] y [16].

Sea C una clase de interés, y X una variable aleatoria condicionante, el diagnóstico estadístico *épsilon* se define de la siguiente manera:

$$\varepsilon = \frac{N_X\big(P(C|X) - P(C)\big)}{\sqrt{N_X P(C)(1 - P(C))}}$$

En esta fórmula, la hipótesis nula $P(C) = N_C/N$ es la probabilidad asociada a que algún elemento de la población pertenezca a la clase de interés, donde N_C es el número de elementos de la población que pertenecen a dicha clase y N es el número de elementos dentro de la población total. $P(C|X) = N_{CX}/N_X$ es la probabilidad de que algún elemento pertenezca a la clase de interés, dado que se sabe que éste cumple con alguna condición $X = x$. En este sentido, el numerador es una medida de la diferencia entre el número de elementos que pertenecen a la clase C y que además cumplen con la condición $X = x$, y el número de elementos que pertenecen a la clase C independientemente de las observaciones realizadas (i.e. con respecto al número de elementos que se esperarían si la hipótesis nula fuera válida). Por otro lado, el denominador es la desviación estándar de la distribución binomial. Si aproximamos a la distribución binomial con la distribución normal, un cociente mayor a 1.96 corresponde al intervalo de confianza del 95% de que las observaciones hechas no son consistentes con la hipótesis nula. Dicho de otro modo, un cociente mayor a 1.96 indican que el número de elementos de la población que cumplen la condición $X = x$ y que además pertenecen a la clase de interés C es inusual con respecto al número de elementos esperado. En este sentido, cuando $\varepsilon > 2$ (pues $1.96 < 2$) decimos que C y $X = x$ están positivamente correlacionadas de forma estadísticamente

significativa. Cuando $\varepsilon < -2$, decimos que C y $X = x$ están negativamente correlacionadas de forma estadísticamente significativa.

En nuestro contexto la clase de interés C será cualquiera de las clases $O, H, A\ y\ S$, y la variable condicionante X será nuestro patrón de historial de hábitos $\boldsymbol{X}$ definido en la sección 4.1.1.1.

En este sentido, nuestro "Análisis de Épsilons" consistirá en calcular ε para todos los 81 patrones de historias $\boldsymbol{X}$ posibles, y todas las clases de interés $O, H, A\ y\ S$. Diremos pues que una persona perteneciente a alguna de estas clases C es "propensa" a mostrar un historial de hábitos $\boldsymbol{X}$ si $\varepsilon(C, \boldsymbol{X}) > 2$. En general, por cada combinación de patrón $\boldsymbol{X}$ y clase C analizada, mostraremos sólo los 10 épsilons más grandes, y a partir de ellos definiremos si alguno indica la presencia de alguna relación estadísticamente significativa que pueda ayudar a perfilar las historias a las que están asociadas las personas que pertenecen a la clase de prueba C.

4.1.1 – Resultados del "Análisis de Épislons"

Los resultados de los análisis de épsilons para las clases O, H, A, S se muestran en las subsecciones siguientes.

4.1.1.1 – Clase O

En el contexto de las personas con obesidad (clase O), analizando la variable *Ejercicio,* predominan patrones donde los hábitos de ejercicio reportados tienden a ser malos en los últimos 5 años (tabla 4.1). Es decir, los patrones del tipo BBB***, BB****, B*****. No obstante, quizá lo más notable es que en los tres patrones más correlacionados con obesidad, hay un "A" reportado para hace 10 años. Esto es, tenemos BBBA**, BB*A**, B*BA**. Esto habla de que, en las personas con obesidad, persiste la historia de la pérdida de un hábito de ejercicio que era bueno hace 10 años, y este historial de *pérdida* de un hábito bueno está más relacionado con el desarrollo de obesidad que lo que un historial de preservación de un hábito malo está relacionado con el mismo desarrollo de obesidad. Dicho de otro modo, una persona es más propensa a desarrollar obesidad si ha perdido un buen hábito de ejercicio que a desarrollar obesidad si ha mantenido un mal hábito de ejercicio a lo largo de su vida. Esto es consistente con lo encontrado por Stephens en [5].

feat	cat	class	classcat	epsilon	nx	ncx	nc
ejerHistory	BB*A**	obesity	1	3.570869	219	68	228
ejerHistory	B*BA**	obesity	1	3.559182	108	38	228
ejerHistory	BBBA**	obesity	1	3.55419	94	34	228
ejerHistory	B**A**	obesity	1	3.449071	308	90	228
ejerHistory	B*****	obesity	1	3.170324	651	171	228
ejerHistory	BB****	obesity	1	3.161199	509	137	228
ejerHistory	*BB***	obesity	1	3.118233	399	110	228
ejerHistory	BBB***	obesity	1	3.02177	353	98	228
ejerHistory	*B****	obesity	1	2.993859	608	159	228
ejerHistory	*BBA**	obesity	1	2.980287	106	35	228

Tabla 4.1. Tabla de los 10 patrones $\boldsymbol{X}$ con los épsilons mayores para los patrones de historias asociados a la variable *Ejercicio*, para la clase O

De igual modo, para *Condición Física*, predominan patrones similares con respecto a los *Ejercicio*, con la diferencia de que el patrón de pérdida de hábitos que eran buenos hace 10 años, pero los bloques "B" se extienden hasta los 10 años. Es decir, predominan historias B*****, BB****, BBB*** y BBBB** (tabla 4.2).

feat	cat	class	classcat	epsilon	nx	ncx	nc
condiHistory	B*B***	obesity	1	5.252381	334	110	228
condiHistory	BBB***	obesity	1	5.199267	317	105	228
condiHistory	*BB***	obesity	1	4.597396	345	108	228
condiHistory	BB****	obesity	1	3.99147	563	158	228
condiHistory	B*****	obesity	1	3.96838	662	182	228
condiHistory	BBBB**	obesity	1	3.853681	157	53	228
condiHistory	**B***	obesity	1	3.791482	409	118	228
condiHistory	B*BB**	obesity	1	3.662919	168	55	228
condiHistory	B*BA**	obesity	1	3.626665	165	54	228
condiHistory	BB*B**	obesity	1	3.474123	168	54	228

Tabla 4.2. Tabla de los 10 patrones $\boldsymbol{X}$ con los épsilons mayores para los patrones de historias asociados a la variable *Condición Física*, para la clase O

Para *Salud*, las personas con obesidad parecen mostrar el punto en común de la pérdida de una buena condición de esta variable en los últimos 5 años. Es decir, dentro de los patrones más significativos, destacan los patrones B**A**, B*A***, B*AA** (tabla 4.3). Esto parece establecer una relación de causalidad *Pérdida de buenos hábitos relacionados con Salud => Obesidad*, y establece un fenómeno similar a lo ocurrido con *Ejercicio*: una persona parece más propensa a desarrollar obesidad si ha perdido una buena condición de salud que a desarrollar obesidad si ha mantenido una mala condición de salud a lo largo de su vida.

feat	cat	class	classcat	epsilon	nx	ncx	nc
saludHistory	B*****	obesity	1	6.123964	441	146	228
saludHistory	B**A**	obesity	1	5.399182	341	113	228
saludHistory	BB****	obesity	1	5.374599	338	112	228
saludHistory	B*A***	obesity	1	4.829568	275	91	228
saludHistory	B*AA**	obesity	1	4.5797	259	85	228
saludHistory	BB*A**	obesity	1	4.438881	255	83	228
saludHistory	BBB***	obesity	1	3.961761	155	53	228
saludHistory	B*B***	obesity	1	3.765409	166	55	228
saludHistory	BBA***	obesity	1	3.658205	183	59	228
saludHistory	BAAA**	obesity	1	3.411579	82	30	228

Tabla 4.3. Tabla de los 10 patrones $\boldsymbol{X}$ con los épsilons mayores para los patrones de historias asociados a la variable *Salud*, para la clase *O*

Para *Estrés*, los patrones para los que $\varepsilon > 2$ tienen la forma AB*A**, ABAA**, A*BA** (tabla 4.4), lo que revela una narrativa constante: la obesidad parece estar relacionada con una situación inestable de estrés en el tiempo reciente (hace $1 - 5$ años), sin que una situación mala sea necesariamente sostenida (pues, en todos los casos, todos estos patrones muestran una situación de estrés buena en el tiempo actual). Si relacionamos esta narrativa a las historias de los hábitos de *Ejercicio* y *Salud*, en las cuales estamos viendo una pérdida de buenos hábitos en los últimos $5 - 10$ años, la condición de estrés mala podría expresarse como un *resultado* de esa pérdida de buenos hábitos. Es decir, el estrés es otro subproducto de la pérdida de buenos hábitos (como parece serlo la obesidad), o bien, hay una causalidad *Pérdida de buenos hábitos en los últimos 5 – 10 años => Situación inestable de Estrés en los últimos 1 – 5 años.* No obstante, esto no explica directamente por qué la condición de estrés asociada a estos patrones tiende a ser buena en el tiempo actual; una explicación posible es que la gente "aprenda" a vivir con esta situación de estrés y se vuelva tolerante a ella.

feat	cat	class	classcat	epsilon	nx	ncx	nc
estresHistory	AB*A**	obesity	1	3.857096	4	4	228
estresHistory	ABAA**	obesity	1	3.340344	3	3	228
estresHistory	AB****	obesity	1	2.42511	92	29	228
estresHistory	A*BA**	obesity	1	2.202608	13	6	228
estresHistory	ABBA**	obesity	1	1.928548	1	1	228
estresHistory	ABA***	obesity	1	1.782751	15	6	228
estresHistory	AABA**	obesity	1	1.735823	12	5	228
estresHistory	ABB***	obesity	1	1.654973	76	22	228
estresHistory	*BAA**	obesity	1	1.592872	34	11	228
estresHistory	A**A**	obesity	1	1.468204	120	32	228

Tabla 4.4. Tabla de los 10 patrones $\boldsymbol{X}$ con los épsilons mayores para los patrones de historias asociados a la variable *Estrés*, para la clase O

Al analizar *Peso*, se revela una narrativa un tanto distinta: las personas con obesidad tienen patrones sostenidos de peso malo, sin haber reportado ninguna instancia de pesos "buenos" en los últimos diez años. Es decir, persisten los patrones BB****, BBB*** y BBBB** (tabla 4.5). Esto puede contraponerse a la hipótesis de causalidad de que *Pérdida de buenos hábitos de Ejercicio y Pérdida de buenos hábitos relacionados con Salud => Obesidad*, pues parecería que la condición de peso mala precede a la pérdida de buenos hábitos y, en todo caso, podría sugerir la dirección de causalidad *Obesidad sostenida => Pérdida de buenos hábitos de Ejercicio y Pérdida de buenos hábitos relacionados con Salud* en el mediano plazo. El reporte de peso, no obstante, tiene una gran limitante cuando la queremos asociar con la obesidad real: *Peso* no es un reflejo del *IMC*, sino de la autopercepción del peso, y las personas tienden a considerar su peso como peor de lo que en realidad es.

feat	cat	class	classcat	epsilon	nx	ncx	nc
pesoHistory	BB****	obesity	1	10.98473	483	201	228
pesoHistory	BBB***	obesity	1	9.945115	285	129	228
pesoHistory	B*****	obesity	1	9.451471	594	220	228
pesoHistory	*B****	obesity	1	9.420686	537	203	228
pesoHistory	B*B***	obesity	1	9.255092	306	131	228
pesoHistory	*BB***	obesity	1	8.81578	316	131	228
pesoHistory	BBBB**	obesity	1	8.648729	147	74	228
pesoHistory	BB*B**	obesity	1	8.423762	153	75	228
pesoHistory	B*BB**	obesity	1	8.157393	154	74	228
pesoHistory	*BBB**	obesity	1	7.950481	160	75	228

Tabla 4.5. Tabla de los 10 patrones $\boldsymbol{X}$ con los épsilons mayores para los patrones de historias asociados a la variable *Peso*, para la clase O

4.1.1.2 – Clase H

Para la división de la población dada por la clase *H*, en términos de la variable *Ejercicio*, las personas con alto grado de estudios mostraron estar asociadas a patrones sostenidamente buenos en los últimos 10 años. Es decir, sobresalen los patrones AA****, AAA*** y AAAA** (tabla 4.6). En cambio, las personas sin alto grado de estudios mostraron estar más asociados a patrones de hábitos sostenidamente malos (tabla 4.7).

feat	cat	class	classcat	epsilon	nx	ncx	nc
ejerHistory	A**A**	hdegree	1	6.030113	285	219	638
ejerHistory	A*****	hdegree	1	5.693315	424	309	638
ejerHistory	A*A***	hdegree	1	5.534243	306	229	638
ejerHistory	AA****	hdegree	1	5.528393	324	241	638
ejerHistory	*A****	hdegree	1	5.312835	465	332	638
ejerHistory	*A*A**	hdegree	1	5.18821	321	236	638
ejerHistory	AA*A**	hdegree	1	5.180005	233	177	638
ejerHistory	*AA***	hdegree	1	5.061124	356	258	638
ejerHistory	A*AA**	hdegree	1	5.032842	251	188	638
ejerHistory	AAA***	hdegree	1	4.989124	253	189	638

Tabla 4.6. Tabla de los 10 patrones $\boldsymbol{X}$ con los épsilons mayores para los patrones de historias asociados a la variable *Ejercicio*, para la clase *H*

feat	cat	class	classcat	epsilon	nx	ncx	nc
ejerHistory	BB*B**	hdegree	0	6.749172	287	173	438
ejerHistory	BBBB**	hdegree	0	6.65125	257	157	438
ejerHistory	*BBB**	hdegree	0	6.44868	290	172	438
ejerHistory	*B*B**	hdegree	0	6.407924	333	193	438
ejerHistory	B*BB**	hdegree	0	6.328181	282	167	438
ejerHistory	BBB***	hdegree	0	6.208445	353	201	438
ejerHistory	B**B**	hdegree	0	6.191506	339	194	438
ejerHistory	B*B***	hdegree	0	6.007127	392	218	438
ejerHistory	*BB***	hdegree	0	5.76574	399	219	438
ejerHistory	BB****	hdegree	0	5.756514	509	271	438

Tabla 4.7. Tabla de los 10 patrones $\boldsymbol{X}$ con los épsilons mayores para los patrones de historias asociados a la variable *Ejercicio*, para la clase *NH*

Para *Condición Física y Salud*, se observan patrones similares para las poblaciones *H* y *NH*: para la población *H*, sobresalen historiales de preservación de buenos hábitos, mientras que para la población *NH*, se observa la pérdida de hábitos que eran buenos hace 5 años o más (tablas 4.8 – 4.11)

feat	cat	class	classcat	epsilon	nx	ncx	nc
condiHistory	A*****	hdegree	1	4.984945	412	294	638
condiHistory	AA****	hdegree	1	4.704237	358	256	638
condiHistory	A**A**	hdegree	1	4.502818	344	245	638
condiHistory	*A****	hdegree	1	4.477774	457	318	638
condiHistory	A*A***	hdegree	1	4.12252	337	237	638
condiHistory	AA*A**	hdegree	1	4.1086	311	220	638
condiHistory	AAA***	hdegree	1	3.877758	311	218	638
condiHistory	A*AA**	hdegree	1	3.695739	315	219	638
condiHistory	*A*A**	hdegree	1	3.628358	394	269	638
condiHistory	*AA***	hdegree	1	3.4885	393	267	638

Tabla 4.8. Tabla de los 10 patrones $\boldsymbol{X}$ con los épsilons mayores para los patrones de historias asociados a la variable *Condición Física*, para la clase *H*

feat	cat	class	classcat	epsilon	nx	ncx	nc
condiHistory	BB****	hdegree	0	4.445665	563	281	438
condiHistory	B*****	hdegree	0	3.838786	662	318	438
condiHistory	*B****	hdegree	0	3.756522	617	297	438
condiHistory	BB*A**	hdegree	0	3.549829	394	195	438
condiHistory	BBA***	hdegree	0	3.486125	246	127	438
condiHistory	BBAA**	hdegree	0	3.364652	235	121	438
condiHistory	B**A**	hdegree	0	3.246159	477	229	438
condiHistory	B*A***	hdegree	0	2.976457	328	160	438
condiHistory	B*AA**	hdegree	0	2.880468	312	152	438
condiHistory	*B*A**	hdegree	0	2.874723	427	203	438

Tabla 4.9. Tabla de los 10 patrones $\boldsymbol{X}$ con los épsilons mayores para los patrones de historias asociados a la variable *Condición Física*, para la clase *NH*

feat	cat	class	classcat	epsilon	nx	ncx	nc
saludHistory	AA****	hdegree	1	6.204964	543	393	638
saludHistory	A*****	hdegree	1	6.065714	631	449	638
saludHistory	AA*A**	hdegree	1	5.960881	494	358	638
saludHistory	AAA***	hdegree	1	5.828194	481	348	638
saludHistory	A**A**	hdegree	1	5.759124	560	399	638
saludHistory	A*A***	hdegree	1	5.731073	536	383	638
saludHistory	AAAA**	hdegree	1	5.525143	457	329	638
saludHistory	A*AA**	hdegree	1	5.399429	508	361	638
saludHistory	*A****	hdegree	1	4.430843	647	439	638
saludHistory	*A*A**	hdegree	1	4.180369	581	394	638

Tabla 4.10. Tabla de los 10 patrones $\boldsymbol{X}$ con los épsilons mayores para los patrones de historias asociados a la variable *Salud*, para la clase *H*

feat	cat	class	classcat	epsilon	nx	ncx	nc
saludHistory	B*****	hdegree	0	7.025779	441	252	438
saludHistory	B**A**	hdegree	0	6.634718	341	199	438
saludHistory	BB****	hdegree	0	6.356441	338	195	438
saludHistory	B*AA**	hdegree	0	5.89016	259	152	438
saludHistory	BB*A**	hdegree	0	5.761327	255	149	438
saludHistory	B*A***	hdegree	0	5.53053	275	157	438
saludHistory	*B****	hdegree	0	5.337306	427	228	438
saludHistory	*B*A**	hdegree	0	4.98259	322	175	438
saludHistory	BBAA**	hdegree	0	4.735178	177	103	438
saludHistory	BBB***	hdegree	0	4.562304	155	91	438

Tabla 4.11. Tabla de los 10 patrones $\boldsymbol{X}$ con los épsilons mayores para los patrones de historias asociados a la variable *Salud*, para la clase *NH*

Para *Estrés*, la población *H* muestra historiales de preservación de condiciones buenas en el corto plazo (patrones AA****, AAA***) y de cambio en la condición de estrés de mala a buena en los últimos diez años (AA*B, A*AB). Por otro lado, la población *NH* muestra condiciones de estrés sostenidamente malas durante los últimos 10 años, es decir, patrones BBB***, *BBB**, BBBB**.

feat	cat	class	classcat	epsilon	nx	ncx	nc
estresHistory	*AA***	hdegree	1	4.534437	275	200	638
estresHistory	AA****	hdegree	1	4.254029	414	288	638
estresHistory	**A***	hdegree	1	4.242096	360	253	638
estresHistory	A*A***	hdegree	1	4.161428	237	172	638
estresHistory	AAA***	hdegree	1	4.148633	222	162	638
estresHistory	*A****	hdegree	1	4.132412	552	375	638
estresHistory	**AB**	hdegree	1	4.09216	206	151	638
estresHistory	A*AB**	hdegree	1	3.734374	130	98	638
estresHistory	AA*B**	hdegree	1	3.691205	298	208	638
estresHistory	*AAB**	hdegree	1	3.396618	157	114	638

Tabla 4.12. Tabla de los 10 patrones $\boldsymbol{X}$ con los épsilons mayores para los patrones de historias asociados a la variable *Estrés*, para la clase *H*

feat	cat	class	classcat	epsilon	nx	ncx	nc
estresHistory	*BBB**	hdegree	0	4.835047	395	208	438
estresHistory	*BB***	hdegree	0	4.677406	435	225	438
estresHistory	BBBB**	hdegree	0	4.119456	319	166	438
estresHistory	*B****	hdegree	0	4.09495	521	258	438
estresHistory	BBB***	hdegree	0	4.009579	358	183	438
estresHistory	B*BB**	hdegree	0	3.969011	396	200	438
estresHistory	*B*B**	hdegree	0	3.792864	444	220	438
estresHistory	B*B***	hdegree	0	3.686622	442	218	438
estresHistory	BB****	hdegree	0	3.52003	428	210	438
estresHistory	BB*B**	hdegree	0	3.412672	357	177	438

Tabla 4.13. Tabla de los 10 patrones $\boldsymbol{X}$ con los épsilons mayores para los patrones de historias asociados a la variable *Estrés*, para la clase *NH*

Para la variable *Peso*, la población *H* muestra patrones predominantemente buenos (AA****, AAA***, AAAA**), pero la población *NH* muestra un fenómeno interesante: éstos presentan historiales que sugieren la pérdida de una buena condición de peso en los últimos 5 - 10 años (tablas 4.14 y 4.15). Es decir, aparecen en ellos los patrones B**A**, BB*A**, BBA***. Esto puede explicarse como que, en esta población, la pérdida de una buena condición de peso en los últimos 5 años es producto de la pérdida de buenos hábitos en el mismo plazo.

feat	cat	class	classcat	epsilon	nx	ncx	nc
pesoHistory	A*****	hdegree	1	4.522458	478	332	638
pesoHistory	AA*A**	hdegree	1	4.451466	377	266	638
pesoHistory	AA****	hdegree	1	4.443251	425	297	638
pesoHistory	A**A**	hdegree	1	4.354067	414	289	638
pesoHistory	A*AA**	hdegree	1	4.329056	368	259	638
pesoHistory	A*A***	hdegree	1	4.220557	399	278	638
pesoHistory	AAAA**	hdegree	1	4.218077	348	245	638
pesoHistory	AAA***	hdegree	1	3.994766	376	261	638
pesoHistory	*AAA**	hdegree	1	3.439218	430	290	638
pesoHistory	*AA***	hdegree	1	3.271104	466	311	638

Tabla 4.14. Tabla de los 10 patrones $\boldsymbol{X}$ con los épsilons mayores para los patrones de historias asociados a la variable *Peso*, para la clase *H*

feat	cat	class	classcat	epsilon	nx	ncx	nc
pesoHistory	B*****	hdegree	0	3.858829	594	288	438
pesoHistory	BB****	hdegree	0	3.55543	483	235	438
pesoHistory	B**A**	hdegree	0	3.203293	424	205	438
pesoHistory	*B****	hdegree	0	3.110052	537	254	438
pesoHistory	B*B***	hdegree	0	2.843686	306	149	438
pesoHistory	BB*A**	hdegree	0	2.751677	328	158	438
pesoHistory	BBA***	hdegree	0	2.66186	198	99	438
pesoHistory	BBAA**	hdegree	0	2.607531	190	95	438
pesoHistory	B*AA**	hdegree	0	2.557726	271	131	438
pesoHistory	B*A***	hdegree	0	2.54392	287	138	438

Tabla 4.15. Tabla de los 10 patrones $\boldsymbol{X}$ con los épsilons mayores para los patrones de historias asociados a la variable *Peso*, para la clase *NH*

4.1.1.3 – Clase A

En el contexto de la división de la población dada por la clase A, para la variable *Ejercicio*, la población de académicos muestra estar asociada a patrones de preservación de hábitos buenos en los últimos 10 años, es decir, patrones del tipo AAA*** y AAAA** (tabla 4.16). La población *NA*, por otro lado, está asociada a patrones de preservación de hábitos predominantemente malos en el mismo plazo, como *BBB** o BBBB** (tabla 4.17).

feat	cat	class	classcat	epsilon	nx	ncx	nc
ejerHistory	AAA***	academic	1	4.158249	253	106	322
ejerHistory	AAAA**	academic	1	4.037504	211	90	322
ejerHistory	AA*A**	academic	1	3.75867	233	96	322
ejerHistory	A*AA**	academic	1	3.705936	251	102	322
ejerHistory	*AAA**	academic	1	3.668827	286	114	322
ejerHistory	A**A**	academic	1	3.584615	285	113	322
ejerHistory	*A*A**	academic	1	3.527155	321	125	322
ejerHistory	A*A***	academic	1	3.423926	306	119	322
ejerHistory	*AA***	academic	1	3.410165	356	136	322
ejerHistory	AA****	academic	1	3.401872	324	125	322

Tabla 4.16. Tabla de los 10 patrones $\boldsymbol{X}$ con los épsilons mayores para los patrones de historias asociados a la variable *Ejercicio*, para la clase A

feat	cat	class	classcat	epsilon	nx	ncx	nc
ejerHistory	*B*B**	academic	0	2.830428	333	257	754
ejerHistory	*BBB**	academic	0	2.793476	290	225	754
ejerHistory	*BB***	academic	0	2.777174	399	305	754
ejerHistory	**BB**	academic	0	2.624461	364	278	754
ejerHistory	**B***	academic	0	2.547718	509	383	754
ejerHistory	BBB***	academic	0	2.514892	353	269	754
ejerHistory	AB*B**	academic	0	2.500379	46	40	754
ejerHistory	*B****	academic	0	2.475124	608	454	754
ejerHistory	***B**	academic	0	2.319229	475	356	754
ejerHistory	BBBB**	academic	0	2.303287	257	197	754

Tabla 4.17. Tabla de los 10 patrones $\boldsymbol{X}$ con los épsilons mayores para los patrones de historias asociados a la variable *Ejercicio*, para la clase *NA*

Para *Condición Física*, la población de académicos *A* muestra estar asociada también a patrones de preservación de hábitos buenos en los últimos diez años (tabla 4.18). Por otro lado, la población *NA* muestra patrones de *Condición Física* donde se observa la pérdida de buenos hábitos en el plazo de los últimos 1-5 años. Es decir, predominan aquí los patrones B**A**, BB*A**, B*AA (tabla 4.19). Como se analizó anteriormente, la clase de las personas sin alto grado de estudios (*NH*) mostraron una pérdida de buenos hábitos de condición física similar, en el mismo plazo que las personas pertenecientes a *NA*. Sabiendo que la población *NH* es una subpoblación de la población *NA* (pues todas las personas sin alto grado de estudios son no-académicos), estos hechos crean una generalización interesante: la pérdida de buenos hábitos de condición física presente en la población *NH* no está asociado necesariamente al grado de estudios que esta población tiene, sino a su tipo de trabajo (*NA*). Esto descarta pues la causalidad *No-Alto Grado de Estudios => Propensión a perder de buenos hábitos de Condición Física.*

feat	cat	class	classcat	epsilon	nx	ncx	nc
condiHistory	AA****	academic	1	3.677791	358	139	322
condiHistory	AAA***	academic	1	3.458665	311	121	322
condiHistory	AA*A**	academic	1	3.334837	311	120	322
condiHistory	A*****	academic	1	3.088358	412	152	322
condiHistory	A**A**	academic	1	2.950034	344	128	322
condiHistory	AAAA**	academic	1	2.933579	291	110	322
condiHistory	A*A***	academic	1	2.872836	337	125	322
condiHistory	AAAB**	academic	1	2.601891	17	10	322
condiHistory	*AA***	academic	1	2.466602	393	140	322
condiHistory	A*AA**	academic	1	2.42808	315	114	322

Tabla 4.18. Tabla de los 10 patrones $\boldsymbol{X}$ con los épsilons mayores para los patrones de historias asociados a la variable *Condición Física*, para la clase *A*

feat	cat	class	classcat	epsilon	nx	ncx	nc
condiHistory	B**A**	academic	0	3.074107	477	365	754
condiHistory	*B*A**	academic	0	2.830325	427	326	754
condiHistory	BB*A**	academic	0	2.740142	394	301	754
condiHistory	B*****	academic	0	2.385595	662	492	754
condiHistory	*BAA**	academic	0	2.375589	259	199	754
condiHistory	B*A***	academic	0	2.309775	328	249	754
condiHistory	*BA***	academic	0	2.303604	272	208	754
condiHistory	B*AA**	academic	0	2.270826	312	237	754
condiHistory	B*BA**	academic	0	2.104183	165	128	754
condiHistory	BBAA**	academic	0	2.040647	235	179	754

Tabla 4.19. Tabla de los 10 patrones $\boldsymbol{X}$ con los épsilons mayores para los patrones de historias asociados a la variable *Condición Física*, para la clase *NA*

Para *Salud*, se observa un fenómeno exactamente igual al de *Condición Física*, donde a través de la observación de la similitud entre los patrones de la población *NA* (tabla 4.21) y lo anteriormente establecido para la población *NH* (tabla 4.11), se determina que la pérdida de buenos hábitos relacionados con *Salud* no es causa del no-alto grado de estudios, sino más bien, del tipo de trabajo de las personas.

feat	cat	class	classcat	epsilon	nx	ncx	nc
saludHistory	B*****	academic	0	3.740635	441	345	754
saludHistory	*B****	academic	0	3.570071	427	333	754
saludHistory	BB****	academic	0	3.2247	338	264	754
saludHistory	B*A***	academic	0	3.199329	275	217	754
saludHistory	B**A**	academic	0	3.080139	341	265	754
saludHistory	B*AA**	academic	0	3.05404	259	204	754
saludHistory	*BA***	academic	0	2.898851	239	188	754
saludHistory	*BAA**	academic	0	2.673931	229	179	754
saludHistory	*B*B**	academic	0	2.647239	105	86	754
saludHistory	*B*A**	academic	0	2.599466	322	247	754

Tabla 4.20. Tabla de los 10 patrones $\boldsymbol{X}$ con los épsilons mayores para los patrones de historias asociados a la variable *Salud*, para la clase *NA*

Para *Estrés*, la población *A* muestra estar asociada a patrones de preservación de una condición buena de estrés en los últimos 10 años. Es decir, predominan los patrones AAA*** y AAAA** (tabla 4.21). Para la población *NA*, los patrones de historias de estrés más representativos revelan un cambio en la situación de estrés de malo a bueno en el plazo de $1 - 5$ años (tabla 4.22). Esto contrasta con los resultados de la población *NH*,

quienes mostraron historiales sostenidos de malas condiciones de estrés en los últimos 10 años. En este sentido, podemos concluir que sí hay una asociación de causalidad fuerte entre el no-alto grado de estudios y la condición de estrés reportado de una persona, independientemente del tipo de trabajo que ésta tenga.

feat	cat	class	classcat	epsilon	nx	ncx	nc
estresHistory	AAAA**	academic	1	4.556547	103	52	322
estresHistory	AAA***	academic	1	4.333122	222	96	322
estresHistory	*AA***	academic	1	4.306647	275	115	322
estresHistory	A*AA**	academic	1	4.301186	106	52	322
estresHistory	*AAA**	academic	1	4.158824	118	56	322
estresHistory	A*A***	academic	1	3.840717	237	98	322
estresHistory	AA*A**	academic	1	3.707611	116	53	322
estresHistory	**A***	academic	1	3.483584	360	138	322
estresHistory	*A*A**	academic	1	3.476651	138	60	322
estresHistory	**AA**	academic	1	3.456218	152	65	322

Tabla 4.21. Tabla de los 10 patrones $\boldsymbol{X}$ con los épsilons mayores para los patrones de historias asociados a la variable *Estrés*, para la clase *A*

feat	cat	class	classcat	epsilon	nx	ncx	nc
estresHistory	AB****	academic	0	3.53607	92	80	754
estresHistory	AB*B**	academic	0	3.234533	86	74	754
estresHistory	ABB***	academic	0	3.192129	76	66	754
estresHistory	ABBB**	academic	0	3.13788	75	65	754
estresHistory	A*B***	academic	0	3.060601	267	210	754
estresHistory	A*BB**	academic	0	2.741914	254	198	754
estresHistory	**BB**	academic	0	2.551863	651	486	754
estresHistory	**B***	academic	0	2.3334	710	526	754
estresHistory	*AB***	academic	0	2.277542	275	210	754
estresHistory	*ABB**	academic	0	2.130455	256	195	754

Tabla 4.22. Tabla de los 10 patrones $\boldsymbol{X}$ con los épsilons mayores para los patrones de historias asociados a la variable *Estrés*, para la clase *NA*

Para *Peso*, no se encontró ninguna relación estadísticamente significa entre la pertenencia o no-pertenencia a la clase *A* con respecto al historial de peso de los últimos 10 años (i.e. todos los épsilon calculados resultaron ser < 2).

4.1.1.3 – Clase S

Según la división dada por la clase *S*, para la variable *Ejercicio*, se observa que las mujeres muestran estar muy asociadas con el cambio de un mal hábito a uno bueno en los últimos 5 años (patrones *ABB**, AA***), mientras que los hombres mostraron estar asociados a historiales de preservación de buenos hábitos de ejercicio en los últimos diez años (tablas 4.23 y 4.24).

ejerHistory	***B**	id_sexo	F	2.812478	475	334	690
ejerHistory	**BB**	id_sexo	F	2.57683	364	257	690
ejerHistory	*A*B**	id_sexo	F	2.439096	142	105	690
ejerHistory	*ABB**	id_sexo	F	2.313774	74	57	690
ejerHistory	A**B**	id_sexo	F	2.286287	136	100	690
ejerHistory	*AB***	id_sexo	F	2.21713	109	81	690
ejerHistory	AA*B**	id_sexo	F	2.040859	90	67	690
ejerHistory	B*BB**	id_sexo	F	2.006812	282	197	690
ejerHistory	**B***	id_sexo	F	1.903408	509	347	690
ejerHistory	B**B**	id_sexo	F	1.881065	339	234	690

Tabla 4.23. Tabla de los 10 patrones ***X*** con los épsilons mayores para los patrones de historias asociados a la variable *Ejercicio*, para la categoría *F*

feat	cat	class	classcat	epsilon	nx	ncx	nc
ejerHistory	**AA**	id_sexo	M	2.671375	451	189	386
ejerHistory	***A**	id_sexo	M	2.591625	593	243	386
ejerHistory	AAAA**	id_sexo	M	2.484088	211	93	386
ejerHistory	A*AA**	id_sexo	M	2.363181	251	108	386
ejerHistory	AA*A**	id_sexo	M	2.242042	233	100	386
ejerHistory	*AAA**	id_sexo	M	2.145346	286	120	386
ejerHistory	A**A**	id_sexo	M	2.06991	285	119	386
ejerHistory	BB*A**	id_sexo	M	1.893076	219	92	386
ejerHistory	*A*A**	id_sexo	M	1.843969	321	131	386
ejerHistory	**A***	id_sexo	M	1.781837	565	223	386

Tabla 4.24. Tabla de los 10 patrones ***X*** con los épsilons mayores para los patrones de historias asociados a la variable *Ejercicio*, para la categoría *M*

Para *Condición Física* y *Salud*, las mujeres mostraron estar relacionadas a la preservación de malos hábitos asociados a condición física y salud durante los últimos 10 años, mientras que los hombres mostraron estar asociados a la preservación de buenos hábitos en el mismo plazo (tablas 4.25 – 4.28).

feat	cat	class	classcat	epsilon	nx	ncx	nc
condiHistory	BBB***	id_sexo	F	3.363092	317	232	690
condiHistory	BBBB**	id_sexo	F	3.048644	157	119	690
condiHistory	*BB***	id_sexo	F	2.89199	345	247	690
condiHistory	B*B***	id_sexo	F	2.831293	334	239	690
condiHistory	BB****	id_sexo	F	2.809059	563	393	690
condiHistory	B*****	id_sexo	F	2.794303	662	459	690
condiHistory	BB*B**	id_sexo	F	2.77762	168	125	690
condiHistory	*B****	id_sexo	F	2.630586	617	427	690
condiHistory	B*BB**	id_sexo	F	2.455907	168	123	690
condiHistory	*BBB**	id_sexo	F	2.329242	175	127	690

Tabla 4.25. Tabla de los 10 patrones ***X*** con los épsilons mayores para los patrones de historias asociados a la variable *Condición Física*, para la categoría *F*

feat	cat	class	classcat	epsilon	nx	ncx	nc
condiHistory	AAAA**	id_sexo	M	3.740932	291	135	386
condiHistory	AAA***	id_sexo	M	3.716216	311	143	386
condiHistory	AA****	id_sexo	M	3.699441	358	162	386
condiHistory	AA*A**	id_sexo	M	3.597989	311	142	386
condiHistory	A*****	id_sexo	M	3.513023	412	182	386
condiHistory	A*AA**	id_sexo	M	3.406504	315	142	386
condiHistory	A*A***	id_sexo	M	3.30568	337	150	386
condiHistory	A**A**	id_sexo	M	3.214414	344	152	386
condiHistory	*A****	id_sexo	M	3.029034	457	195	386
condiHistory	*AAA**	id_sexo	M	2.715512	368	157	386

Tabla 4.26. Tabla de los 10 patrones ***X*** con los épsilons mayores para los patrones de historias asociados a la variable *Condición Física*, para la categoría *M*

feat	cat	class	classcat	epsilon	nx	ncx	nc
saludHistory	*B****	id_sexo	F	3.852294	427	312	690
saludHistory	B*B***	id_sexo	F	3.81096	166	130	690
saludHistory	BBB***	id_sexo	F	3.617964	155	121	690
saludHistory	BB****	id_sexo	F	3.430845	338	247	690
saludHistory	*BB***	id_sexo	F	3.412589	188	143	690
saludHistory	*B*A**	id_sexo	F	3.196718	322	234	690
saludHistory	**B***	id_sexo	F	3.138353	260	191	690
saludHistory	B*****	id_sexo	F	2.899321	441	312	690
saludHistory	BBBB**	id_sexo	F	2.761562	77	61	690
saludHistory	B*BB**	id_sexo	F	2.760275	84	66	690

Tabla 4.27. Tabla de los 10 patrones ***X*** con los épsilons mayores para los patrones de historias asociados a la variable *Salud*, para la categoría *F*

feat	cat	class	classcat	epsilon	nx	ncx	nc
saludHistory	AAAA**	id_sexo	M	3.711741	457	202	386
saludHistory	*AAA**	id_sexo	M	3.703936	540	235	386
saludHistory	*AA***	id_sexo	M	3.488393	574	246	386
saludHistory	AAA***	id_sexo	M	3.464932	481	209	386
saludHistory	AA*A**	id_sexo	M	3.35679	494	213	386
saludHistory	*A*A**	id_sexo	M	3.250106	581	246	386
saludHistory	AA****	id_sexo	M	3.239504	543	231	386
saludHistory	*A****	id_sexo	M	3.106387	647	270	386
saludHistory	A*AA**	id_sexo	M	2.938131	508	214	386
saludHistory	A*A***	id_sexo	M	2.676231	536	222	386

Tabla 4.28. Tabla de los 10 patrones $\boldsymbol{X}$ con los épsilons mayores para los patrones de historias asociados a la variable *Salud*, para la categoría *M*

Para *Estrés* (tablas 4.29 y 4.30), las mujeres están asociadas a condiciones buenas de estrés en los últimos 5-10 años (AAA***, A*AA**), mientras que los hombres están asociados a reportes de preservación de malas condiciones de estrés en el mismo plazo (patrones BBB***, BBBB**). Ya que los hombres mostraron tener historiales de *Ejercicio, Condición Física* y *Salud* sostenidamente buenos, este resultado al respecto de la variable *Estrés* es interesante, pues esencialmente muestra que una preservación de buenos hábitos y la manutención de malos hábitos en su mayoría es independiente de la condición de *Estrés* de una persona y que, por lo tanto, el historial de *Estrés* es más causa de otros factores externos.

feat	cat	class	classcat	epsilon	nx	ncx	nc
estresHistory	A*****	id_sexo	F	3.199597	506	359	690
estresHistory	AA****	id_sexo	F	3.024559	414	295	690
estresHistory	*A****	id_sexo	F	2.752952	552	385	690
estresHistory	A**A**	id_sexo	F	2.673795	120	91	690
estresHistory	***A**	id_sexo	F	2.50097	207	150	690
estresHistory	AA*A**	id_sexo	F	2.441724	116	87	690
estresHistory	A*B***	id_sexo	F	2.268987	267	189	690
estresHistory	A*AA**	id_sexo	F	2.232852	106	79	690
estresHistory	AAA***	id_sexo	F	2.188456	222	158	690
estresHistory	A*A***	id_sexo	F	2.169674	237	168	690

Tabla 4.29. Tabla de los 10 patrones $\boldsymbol{X}$ con los épsilons mayores para los patrones de historias asociados a la variable *Estrés*, para la categoría *F*

feat	cat	class	classcat	epsilon	nx	ncx	nc
estresHistory	BB*B**	id_sexo	M	4.516632	357	169	386
estresHistory	BBBB**	id_sexo	M	4.268182	319	151	386
estresHistory	*B*B**	id_sexo	M	3.732398	444	197	386
estresHistory	BBB***	id_sexo	M	3.699441	358	162	386
estresHistory	BB****	id_sexo	M	3.674518	428	190	386
estresHistory	B*BB**	id_sexo	M	3.660799	396	177	386
estresHistory	B**B**	id_sexo	M	3.615718	472	207	386
estresHistory	*BBB**	id_sexo	M	3.388348	395	174	386
estresHistory	B*B***	id_sexo	M	3.117791	442	190	386
estresHistory	B*****	id_sexo	M	3.063374	566	238	386

Tabla 4.30. Tabla de los 10 patrones $\boldsymbol{X}$ con los épsilons mayores para los patrones de historias asociados a la variable *Estrés*, para la categoría *M*

Para *Peso*, el único patrón relacionado de forma estadísticamente significativa con las mujeres es el patrón BABA** (tabla 4.31), no obstante, la cantidad de evidencia que resulta en ese patrón es poca. Puede ser que algunas mujeres dentro de esa pequeña cantidad de evidencia hayan estado embarazadas en los últimos 10 años, y por eso hayan visto esa fluctuación entre sus estados de *Peso*.

feat	cat	class	classcat	epsilon	nx	ncx	nc
pesoHistory	BAB***	id_sexo	F	2.517567	21	19	690
pesoHistory	BABA**	id_sexo	F	2.241326	14	13	690
pesoHistory	ABBA**	id_sexo	F	1.566834	17	14	690
pesoHistory	B*B***	id_sexo	F	1.522419	306	209	690
pesoHistory	*B*A**	id_sexo	F	1.412007	365	247	690
pesoHistory	BB****	id_sexo	F	1.353723	483	324	690
pesoHistory	*B****	id_sexo	F	1.317302	537	359	690
pesoHistory	B*****	id_sexo	F	1.290826	594	396	690
pesoHistory	BABB**	id_sexo	F	1.19084	7	6	690
pesoHistory	**BA**	id_sexo	F	1.18977	198	135	690

Tabla 4.31. Tabla de los 10 patrones $\boldsymbol{X}$ con los épsilons mayores para los patrones de historias asociados a la variable *Peso*, para la categoría *F*

Capítulo 5. Modelos de Predicción

Partiendo de los análisis establecidos en el capítulo anterior, hay suficiente evidencia para establecer que ciertos grupos de la población general están más asociados con ciertos historiales de hábitos. Con esto en mente, en el presente capítulo, intentaremos predecir si una persona con un cierto historial de hábitos pertenece o no a un grupo de la población determinado. Específicamente, intentaremos predecir si una persona padece o no de obesidad, dados sus distintos historiales de hábitos. Para hacer esto, construiremos y evaluaremos diversos modelos a partir de la metodología que se describe a continuación

5.1 – Herramientas para la construcción de los modelos de predicción

Formalmente, buscaremos construir un modelo que pueda predecir (o bien, clasificar) si una persona con un historial de hábitos $\boldsymbol{X}$ asociado a alguna variable de *Autoevaluación de Salud* $\boldsymbol{a}$, pertenece o no a la clase *Obesidad*. Para construir dicho modelo, necesitamos distintas herramientas:

1. Una o varias técnicas de aprendizaje supervisado para clasificación binaria. Un algoritmo que pueda encontrar una función capaz de mapear una entrada (i.e. el patrón de historial de hábitos de una persona) a una salida (i.e. *Obesidad = 1* para el caso de un individuo con obesidad, u *Obesidad = 0*, para el caso de un individuo sin obesidad), basado en un conjunto de ejemplos (entrenamiento) de pares entrada-salida.

2. Una serie de técnicas de evaluación de los modelos de clasificación binaria creados a partir de las técnicas elegidas. Una o varias métricas que nos permitan comparar a distintos modelos (creados a partir de técnicas de clasificación binaria distintas) entre sí, para definir cuál de los modelos es el más adecuado para la tarea actual.

Todos estos mecanismos los planteamos a continuación.

5.1.1 – Técnicas de clasificación binaria

En el ánimo de continuar el estudio iniciado por Stephens et. al en [5], planteamos el usar los mismos modelos de clasificación ahí presentados: *Naive Bayes* y *Generalized Naive Bayes*.

5.1.1.1 – *Naive Bayes* (NBA)

El clasificador *Naive Bayes* es un algoritmo de clasificación probabilística basado en la aplicación del Teorema de Bayes, y que asume independencia entre las variables de entrada [17].

Sea C la clase "Obesidad", donde $C = 1$ corresponde a un individuo que es obeso, y $C = 0$ corresponde a uno que no es obeso, y sea $\boldsymbol{X} = (x_1, x_2, \dots, x_n)$ un vector de observaciones tales que x_i corresponde dentro de nuestro contexto al valor de alguna *variable de autoevaluación de salud* a_t en algún tiempo t (donde $a_t \in \{A, B, N\}$ y $t \in \{0, 1, 5, 10, 20, 30\}$[2]), *Naive Bayes* busca determinar la probabilidad *posterior* $P(C|\boldsymbol{X})$ y clasificar a la persona dentro de $C = 0$ ó $C = 1$ con base en dicha probabilidad.

Para realizar esta clasificación basándonos en la estimación de $P(C|\boldsymbol{X})$, partimos del teorema de Bayes, que relaciona a la probabilidad posterior $P(C|\boldsymbol{X})$ con una probabilidad *a priori* $P(C)$ en la ausencia de información:

$$P(C|\boldsymbol{X}) = \frac{P(\boldsymbol{X}|C)P(C)}{P(\boldsymbol{X})}$$

Donde $P(\boldsymbol{X}|C)$ es la probabilidad de observar una evidencia $\boldsymbol{X}$ dada la clase C y $P(\boldsymbol{X})$ es la probabilidad de la evidencia $\boldsymbol{X}$. Con el fin de evitar calcular al factor $P(\boldsymbol{X})$, consideremos el dividir $P(C|\boldsymbol{X})$ entre $P(\bar{C}|\boldsymbol{X})$, de manera que obtengamos:

$$\frac{P(C|\boldsymbol{X})}{P(\bar{C}|\boldsymbol{X})} = \frac{\frac{P(\boldsymbol{X}|C)P(C)}{P(\boldsymbol{X})}}{\frac{P(\boldsymbol{X}|\bar{C})P(\bar{C})}{P(\boldsymbol{X})}} = \frac{P(\boldsymbol{X}|C)P(C)}{P(\boldsymbol{X}|\bar{C})P(\bar{C})}$$

De este modo, definimos a una función de *score* $S(\boldsymbol{C}|\boldsymbol{X})$ como sigue:

$$S(C|\boldsymbol{X}) = \ln\left(\frac{P(C|\boldsymbol{X})}{P(\bar{C}|\boldsymbol{X})}\right) = \ln\left(\frac{P(\boldsymbol{X}|C)P(C)}{P(\boldsymbol{X}|\bar{C})P(\bar{C})}\right) = \ln\left(\frac{P(\boldsymbol{X}|C)}{P(\boldsymbol{X}|\bar{C})}\right) + \ln\left(\frac{P(C)}{P(\overline{C})}\right)$$

[2] Seguimos las mismas convenciones establecidas en el capítulo 4.3.1.1.

Ahora bien, asumiendo independencia de las n variables que componen al vector $\boldsymbol{X}$, tenemos que $P(\boldsymbol{X}|C) = \prod_{i=1}^{n} P(X_i|C)$. Esto es:

$$S(C|\boldsymbol{X}) = \ln\left(\frac{P(\boldsymbol{X}|C)}{P(\boldsymbol{X}|\bar{C})}\right) + \ln\left(\frac{P(C)}{P(\overline{C})}\right) = \ln\left(\frac{\prod_{i=1}^{n} P(X_i|C)}{\prod_{i=1}^{n} P(X_i|\bar{C})}\right) + \ln\left(\frac{P(C)}{P(\overline{C})}\right)$$

Con el fin de simplificar la función anterior, podemos agrupar a los términos $P(X_i|C)$ y $P(X_i|\bar{C})$ de la siguiente forma:

$$\ln\left(\frac{\prod_{i=1}^{n} P(X_i|C)}{\prod_{i=1}^{n} P(X_i|\bar{C})}\right) = \ln\left(\frac{P(X_1|C) \times P(X_2|C) \times \ldots \times P(X_n|C)}{P(X_1|\bar{C}) \times P(X_2|\bar{C}) \times \ldots \times P(X_n|\bar{C})}\right)$$
$$= \ln\left(\frac{P(X_1|C)}{P(X_1|\bar{C})}\right) + \ln\left(\frac{P(X_2|C)}{P(X_2|\bar{C})}\right) + \cdots + \ln\left(\frac{P(X_n|C)}{P(X_n|\bar{C})}\right) = \sum_{i=1}^{n} \ln\left(\frac{P(X_i|C)}{P(X_i|\bar{C})}\right)$$

De manera que podemos entonces expresar a nuestra función de *score* como se muestra a continuación.

$$S(C|\boldsymbol{X}) = \sum_{i=1}^{n} \ln\left(\frac{P(X_i|C)}{P(X_i|\bar{C})}\right) + \ln\left(\frac{P(C)}{P(\overline{C})}\right) = \sum_{i=1}^{n} S(X_i) + \ln\left(\frac{P(C)}{P(\overline{C})}\right)$$

Donde:

$$S(X_i) = \ln\left(\frac{P(X_i|C)}{P(X_i|\bar{C})}\right) = \ln\left(\frac{\frac{N_{CX_i}}{N_C}}{\frac{N_{X_i} - N_{CX_i}}{N - N_C}}\right)$$

Ahora bien, ya que es posible que alguno de los parámetros N sea 0, resultando en una indeterminación (en general es un problema con los clasificadores bayesianos que, al calcular la probabilidad posterior, un factor de 0 en uno de los *likelihoods* resulte en una probabilidad posterior de 0 [18], que en este cálculo del score desemboque en una indeterminación), añadimos un suavizado (*smoothing*) a esta función de score:

$$S(X_i) = \ln\left(\frac{\frac{N_{CX_i}+1}{N_C+2}}{\frac{N_{X_i}-N_{CX_i}+1}{N-N_C+2}}\right)$$

Con esta ecuación, clasificamos a cada uno de nuestros individuos dentro de la clase C si $S(C|\boldsymbol{X}) > 0$ y lo clasificamos dentro de la clase $\bar{C}$ si $S(C|\boldsymbol{X}) < 0$.

De forma general, nuestra función de *score* relaciona las ocurrencias de $X = x$ dada la ocurrencia de una clase C, y las ocurrencias de $X = x$ dada la no-ocurrencia de la clase C. Si el número resultante de este cociente es menor que 1 (i.e., la ocurrencia de $X = x$ está más relacionada con la no-ocurrencia de C que con la ocurrencia de C), $S(X = x)$ resultará en un número negativo. Si el cociente es mayor o igual que 1 (i.e., la ocurrencia de $X = x$ está igual más relacionada con la ocurrencia de C que con la no-ocurrencia de C), $S(X = x)$ resultará en un número positivo. El objetivo del clasificador es pues el de calcular los *scores* para cada variable X_i que compone a $\boldsymbol{X}$, y posteriormente sumar esos scores individuales (pues, al utilizar NB, asumimos independencia entre las variables) para calcular el *score* total que le corresponde a un individuo, según las categorías a las que éste pertenece dentro de cada variable. Ya que $S(X = x)$ "recompensa" a $X = x$ cuando esta categoría está relacionada con la ocurrencia de la clase C, y que "penaliza" a $X = x$ cuando ésta está relacionada con la no-ocurrencia de la clase C, entonces, intuitivamente puede decirse que tendremos un mayor *score* para los individuos cuyas categorías se relacionan más con la ocurrencia que con la no-ocurrencia de la clase *C*. De este argumento, se sigue que aquellos individuos con un mayor score total serán los que, para nuestro clasificador, son más propensos a ser clasificados como obesos. Esto es, nuestra función de *score* no sólo nos ayuda a clasificar a un individuo, sino a establecer un orden entre los individuos, según qué tan fuerte sea el *score* asociado a ellos.

Retomando el contexto de este trabajo, sea pues C la clase "Obesidad", y $\boldsymbol{X} = a_{act}a_1a_5a_{10}a_{20}a_{30}$ el historial de alguna variable de *Autoevaluación de Salud* de alguna persona dentro de la base de datos. Supongamos que la variable en cuestión es *Ejercicio*, y que una persona p tiene el historial de ejercicio $\boldsymbol{X}^p_{Ejercicio} = ABBBBB$. En NBA, el score asociado a este historial se calcula como:

$$\begin{aligned} S\left(C = 1 \middle| \boldsymbol{X}^p_{Ejercicio}\right) &= \sum_{i=1}^{n} S(X_i) \\ &= S(a_{act} = A) + S(a_1 = B) + S(a_5 = B) + S(a_{10} = B) + S(a_{20} = B) + S(a_{30} = B) \end{aligned}$$

5.1.1.2 – *Generalized Naive Bayes* (GNB)

Una clase de generalizaciones del método de NBA antes presentado se puede proponer a partir de la generalización del factor del *likelihood* $P(\boldsymbol{X}|C) = \prod_{i=1}^{n} P(X_i|C)$ como $P(\boldsymbol{X}|C) = \prod_{i=1}^{M_\xi} P(X_\xi|C)$ [19] donde, en lugar de que X_i represente al valor de alguna variable particular, X_ξ representa a una combinación de varias variables $X_i, X_j \ldots$ etcétera. Por ejemplo, dadas cuatro variables binarias $X_1, X_2, X_3, X_4 \in \{A, B\}$, podemos formar las combinaciones de variables $\xi_1 = X_1X_2$ y $\xi_2 = X_3X_4$, donde ξ_1 y ξ_2 pueden tomar $2^2 = 4$ valores posibles: AA, AB, BA, BB. Con esto, el likelihood asociado a X_1, X_2, X_3, X_4, es decir $P(X_1X_2X_3X_4|C)$ puede expresarse como $P(\xi_1\xi_2|C)$. En NBA, $P(X_1X_2X_3X_4|C) = P(X_1|C)P(X_2|C)P(X_3|C)P(X_4|C)$, y usando la agrupación de variables sugerida, $P(\xi_1\xi_2|C) = P(\xi_1|C)P(\xi_2|C) = P(X_1X_2|C)P(X_3X_4|C)$. Es decir, esta generalización nos da una forma de expresar al *likelihood* $P(X_1X_2X_3X_4|C)$ como la factorización $P(X_1X_2|C)P(X_3X_4|C)$. En términos comunes, el poder de esta generalización radica en que esta nueva factorización (o cualquier otra factorización que pueda haberse generado a partir de las variables iniciales) no implica la asunción de independencia entre todas las variables, como sí lo hace la factorización $P(X_1X_2X_3X_4|C) = P(X_1|C)P(X_2|C)P(X_3|C)P(X_4|C)$.

Utilizando la combinación particular $\xi = X_1X_2X_3X_4$ bajo esta generalización, obtenemos la factorización del *likelihood* $P(X_1X_2X_3X_4|C) = P(\xi|C)$, donde ξ puede tomar cualquiera de $2^4 = 16$ valores posibles. Esta combinación no está asumiendo ninguna independencia entre ninguna de las variables, por lo que, en este caso de estudio particular, donde sabemos que la evolución de los hábitos de las personas hacia el futuro no es necesariamente independiente del su historial pasado, teóricamente debería ser más efectivo que NBA a la hora de usar el historial para clasificar a una persona.

Sea pues C la clase "Obesidad", y $\boldsymbol{X} = a_{act}a_1a_5a_{10}a_{20}a_{30}$ el historial de alguna variable de *Autoevaluación de Salud* de alguna persona dentro de la base de datos. Supongamos que la variable en cuestión es *Ejercicio*, y que una persona p tiene el historial de ejercicio $\boldsymbol{X}^p_{Ejercicio} = ABBBBB$. En GNB, usando la agrupación de variables $\xi = a_{act}a_1a_5a_{10}a_{20}a_{30}$, donde ξ toma uno de 2^6 el score asociado a esta persona se calcula como:

$$S\left(C = 1|\boldsymbol{X}^p_{Ejercicio}\right) = S(\xi) = S(a_{act} = A, a_1 = B, a_5 = B, a_{10} = B, a_{20} = B, a_{30} = B)$$

5.1.2 – Técnicas/Métricas de evaluación de modelos de clasificación

Para poder comparar entre el desempeño de los distintos modelos que se crearán a partir de las técnicas antes mencionadas, planteamos el uso de tres métricas distintas: la curva *ROC*, la curva *PRC,* la distribución de verdaderos positivos entre los individuos ordenados según su score. Todas estas métricas se describen a continuación.

5.1.2.1 – Curvas ROC y PRC

La curva ROC, o *Receiver Operating Characteristic*, es un gráfico que ilustra el desempeño de un clasificador binario según varía el umbral de discriminación [20]. Se obtiene graficando a la tasa de verdaderos positivos (*TPR* o *sensibilidad*) contra la tasa de falsos positivos (*FPR*) que arroja el clasificador, usando umbrales de discriminación distintos, donde:

$$TPR = \frac{\#\ de\ Verdaderos\ Positivos\ (TP)}{\#\ de\ Positivos\ (P)} = \frac{TP}{TP + \#\ de\ Falsos\ Negativos\ (FN)}$$

$$FPR = \frac{\#\ de\ Falsos\ Positivos\ (FP)}{\#\ de\ Negativos\ (N)} = \frac{FP}{FP + \#\ de\ Verdaderos\ Negativos\ (TN)}$$

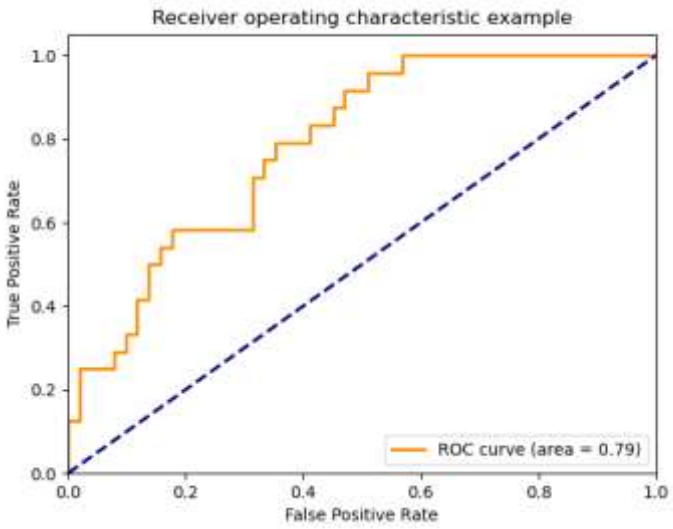


Figura 5.1. Ejemplo de curva ROC [21]

Esencialmente, la curva ROC nos dice qué tanto es capaz nuestro modelo de clasificación de distinguir entre clases. El clasificador ideal clasifica a los sujetos de prueba, usando algún umbral de clasificación determinado, con una tasa de verdaderos positivos igual a 1.0 y una tasa de falsos positivos igual a 0: es decir, es capaz de

identificar a todos los sujetos pertenecientes a la clase C y a aquellos pertenecientes a $\bar{C}$ sin cometer ningún error.

La curva PR, o *Precision Recall Curve* (PRC), por su parte, también es un gráfico que busca ilustrar la calidad de los outputs generados por el modelo de clasificación en cuestión [22]. No obstante, a diferencia de la curva ROC, la curva PRC se genera graficando la precisión (*precision)* contra la exhaustividad (*recall*), donde:

$$Precisión = \frac{TP}{TP + FP}$$

$$Exhaustividad = \frac{TP}{TP + FN}$$

En términos prácticos, la precisión es una medida de la relevancia de los resultados obtenidos, mientras que la exhaustividad es una medida de cuántos resultados obtenidos son relevantes. Un clasificador con alta exhaustividad, pero baja precisión puede regresar muchos resultados positivos, pero muchas de sus predicciones pueden resultar ser incorrectas. En cambio, un clasificador con alta precisión, pero baja exhaustividad puede regresar pocos resultados positivos, pero gran parte de esos resultados son correctos. Un clasificador ideal, desde la perspectiva de la PRC, tiene alta precisión y exhaustividad: el clasificador regresa un gran número de resultados positivos, y gran parte de esos resultados positivos fueron clasificados correctamente. Estas características hacen que la PRC sea particularmente efectiva en la evaluación del desempeño de un clasificador que está trabajando con una clase desbalanceada [23][24]: en un problema como ése, una mejor PRC evaluada sobre la clase minoría puede hablar de un clasificador que es mejor identificando y clasificando correctamente a los individuos de dicha clase, sin importar el desempeño del clasificador sobre la clase mayoría.

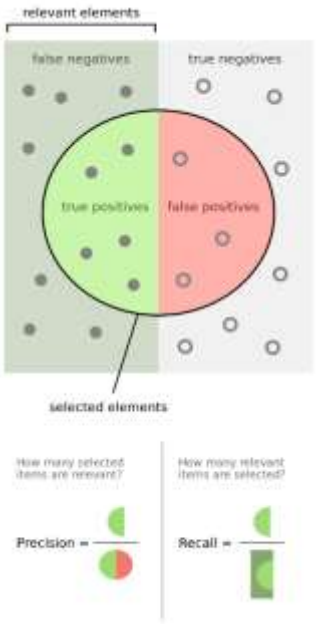


Figura 5.2. Precisión y Exhaustividad [25]

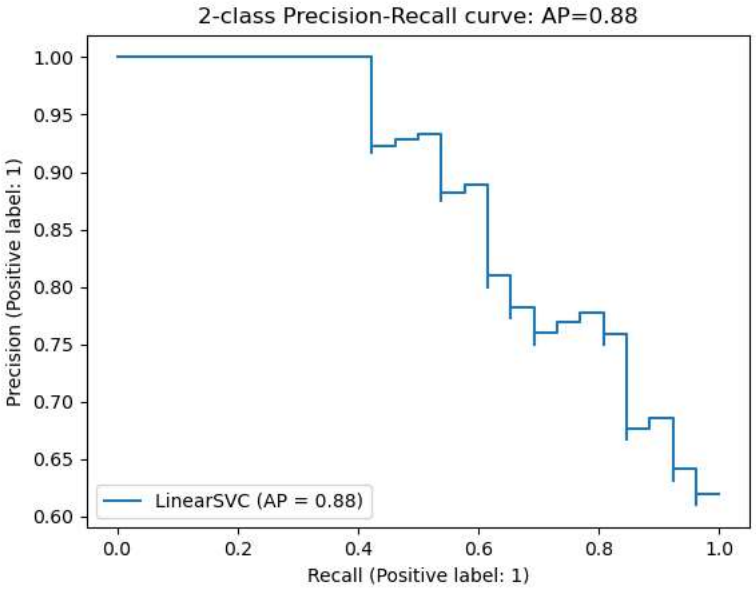


Figura 5.3. Ejemplo de PRC [26]

5.1.2.2 – Ordenamiento de Scores

Los clasificadores presentados en la sección 5.1.1 clasifican a los individuos de la población a partir de un score. Como se discutió anteriormente, un score mayor o menor habla de un mayor o menor sesgo a ser clasificado dentro de una clase C o $\bar{C}$, respectivamente. Este mayor o menor sesgo, en términos comunes, podemos interpretarlo como una "propensión mayor" a ser clasificado dentro de C o $\bar{C}$. Por ejemplo, supongamos que tenemos dos personas p_1 y p_2, cuyos scores de clasificación con respecto a la clase $Obesidad$, dados sus historiales de ejercicio $\boldsymbol{X}^{p_1}$ y $\boldsymbol{X}^{p_2}$, son $S(Obesidad \mid \boldsymbol{X}^{p_1})$ y $S(Obesidad \mid \boldsymbol{X}^{p_2})$, respectivamente. Decimos que, según las evidencias mostradas ($\boldsymbol{X}^{p_1}$ y $\boldsymbol{X}^{p_2}$), p_1 es más propenso a padecer obesidad que p_2 si $S(Obesidad \mid \boldsymbol{X}^{p_1}) > S(Obesidad \mid \boldsymbol{X}^{p_2})$, o viceversa.

En este sentido, el score arrojado por nuestro modelo de clasificación nos brinda una intuición extra al clasificar individuos: dada una clase C, los scores que arroja el modelo no sólo nos ayudan a determinar cuáles individuos deben ser clasificados dentro de C y cuáles dentro de $\bar{C}$, sino que también nos ayudan a establecer un orden entre los individuos clasificados. Los individuos con mayor score son los que más probablemente están asociados con C, y los individuos con menor score son aquellos que están menos probablemente asociados con C.

Con esto en mente, proponemos el usar como criterio de evaluación de nuestros modelos de predicción a la distribución de verdaderos positivos entre los individuos clasificados, según el ordenamiento que sus scores arrojen. Es decir, dada una población P de individuos clasificados en la clase C o $\bar{C}$ por el modelo M, obtendremos nuestra "distribución de positivos, según su orden de scores" como sigue:

1. Obtener los scores de todos los individuos de P.
2. Ordenar a los individuos de P, según sus scores, de mayor a menor.
3. Dividir a los individuos ordenados en deciles: el decil 1 corresponde a los individuos con mayores scores de la población, y el decil 10 corresponde a los individuos con los menores scores de la población.
4. Cuantificar, por decil, a todos los individuos que en realidad pertenecen a la clase C (positivos)

Basándonos en nuestro razonamiento anterior sobre el ordenamiento de scores, un clasificador efectivo debería generar un orden de scores tal que, cuando los individuos son ordenados por deciles bajo este algoritmo, deberíamos ver un desbalance en la distribución de individuos que en realidad pertenecen a la clase C (positivos): los deciles correspondientes a los scores mayores deberían contener a un mayor número de verdaderos positivos de C que los deciles correspondientes a los scores menores, pues hemos establecido que un mayor score implica una mayor asociación a la clase C. Al ordenar a la población asignándoles scores aleatorios, por otro lado, no debería verse una diferencia significativa entre los individuos de uno y otro decil; cada decil debería contener un número más o menos igual a $\frac{\#\ de\ individuos\ positivos\ dentro\ de\ la\ población}{10}$.

Por ejemplo, dada una población de aproximadamente de 650 individuos, donde 130 pertenecen en realidad a la clase C. Si a estos individuos se les asigna un score dado por algún clasificador efectivo, los ordenamos según su score y los separamos en diez deciles bajo este orden, la distribución de los 130 individuos que en realidad corresponden a la clase C debería verse como en la figura mostrada a continuación.

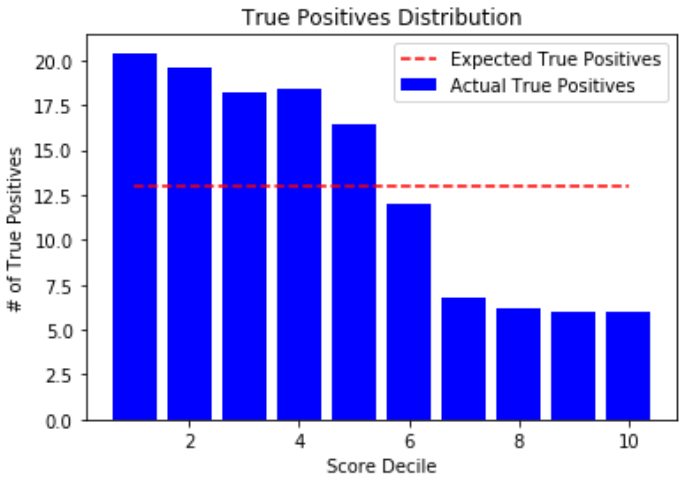


Figura 5.4 Ejemplo de distribución de 130 positivos de una población de 650 individuos, según su orden de scores

Como puede observarse, bajo este ordenamiento de individuos hay una mayor concentración de individuos positivos dentro de los primeros deciles, lo que indica que los scores arrojados por el clasificador sí son significativos a la hora de identificar a aquellos individuos más "propensos" a ser identificados como positivos. Si el clasificador no generara scores de estas características (por ejemplo, si el clasificador asignara los scores aleatoriamente), la distribución de individuos positivos debería verse mucho más uniforme, sin ningún sesgo particular hacia ningún decil (línea roja punteada).

5.2 – Metodología del entrenamiento y evaluación de los modelos de predicción

Como se mencionó en el capítulo 4.1.3, los datos correspondientes a las variables de *Autoevaluación de Salud* no están completos para todas las entradas de la base de datos. Por este motivo, para construir nuestros modelos de clasificación, utilizaremos sólo un subconjunto de la base de datos: aquellas entradas que tienen información completa para cada variable de *Autoevaluación de Salud* en el rango de 0 – 20 años. Es decir, las historias que se analizarán serán todas de la forma $\boldsymbol{X} = a_{act}a_0a_1a_5a_{10}a_{20}$.

Usando las herramientas presentadas anteriormente, en las siguientes subsecciones construiremos los siguientes modelos de predicción:

- Predicción de *Obesidad*, usando los patrones de *Ejercicio* de 0 – 20 años, usando NBA y GNB.
- Predicción de *Obesidad*, usando los patrones de *Salud* de 0 – 20 años, usando NBA y GNB.

- Predicción de *Obesidad*, usando los patrones de *Condición Física* de 0 – 20 años, usando NBA y GNB.
- Predicción de *Obesidad*, usando los patrones de *Estrés* de 0 – 20 años, usando NBA y GNB.

Cada modelo pasará a través de una validación cruzada de 5 particiones. Es decir, la base de datos se reordena aleatoriamente, y luego se realiza el entrenamiento de 5 predictores basados en la función de score dada por el modelo, tomando el 80% de los datos para entrenar y el 20% de los datos para probar la fiabilidad del predictor [27].

Prueba	Entrenamiento	Entrenamiento	Entrenamiento	Entrenamiento
Entrenamiento	Prueba	Entrenamiento	Entrenamiento	Entrenamiento
Entrenamiento	Entrenamiento	Prueba	Entrenamiento	Entrenamiento
Entrenamiento	Entrenamiento	Entrenamiento	Prueba	Entrenamiento
Entrenamiento	Entrenamiento	Entrenamiento	Entrenamiento	Prueba

Figura 5.5. División de la base de datos en cinco particiones para la validación cruzada

De esta validación cruzada de cinco particiones obtendremos:

1. La curva ROC asociada a cada partición.
2. La PRC asociada a cada partición.
3. El promedio de la distribución de los positivos (discutida en la sección 5.1.2.2), según el orden de scores de los datos de prueba de cada partición.

5.3 – Predicción de Obesidad a partir de historiales de Ejercicio

Antes de construir los modelos de predicción de *Obesidad* usando los historiales de variables de *Autoevaluación de Salud*, conviene hacer un análisis preliminar sobre los resultados esperados con respecto al desempeño de cada una de las técnicas utilizadas, NBA y GNB. Esta discusión es de particular interés debido a la asunción que NBA hace al respecto de las variables de entrada que GNB no: la independencia entre las variables de entrada.

NBA asume independencia de variables, no obstante, basándonos en el análisis del capítulo 4 hecho con respecto a la evolución de patrones en los historiales de hábitos de las personas, la evolución de un hábito en el tiempo no es independiente del estado de ese mismo hábito en el pasado. Sobre todo, un hábito no es independiente del tiempo inmediatamente anterior (o en el tiempo de referencia donde la fuerza de preservación de hábitos está más vigente). En este sentido, al no hacer la misma asunción, la intuición

dicta que GNB debería ser mejor predictor que NBA para esta tarea de predicción de *Obesidad*.

No obstante, al construir los primeros modelos de predicción de *Obesidad* basado en los historiales de *Ejercicio*, obtenemos los resultados mostrados a continuación.

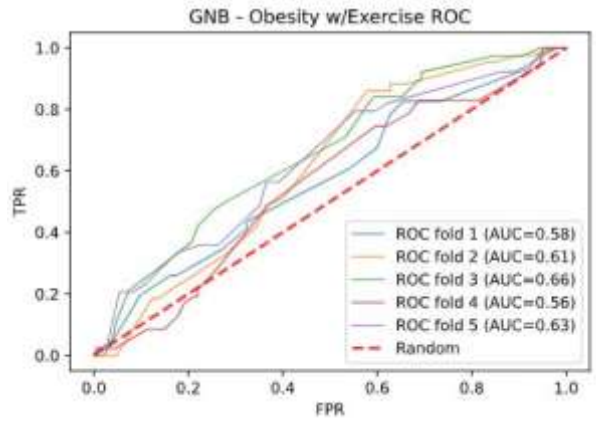
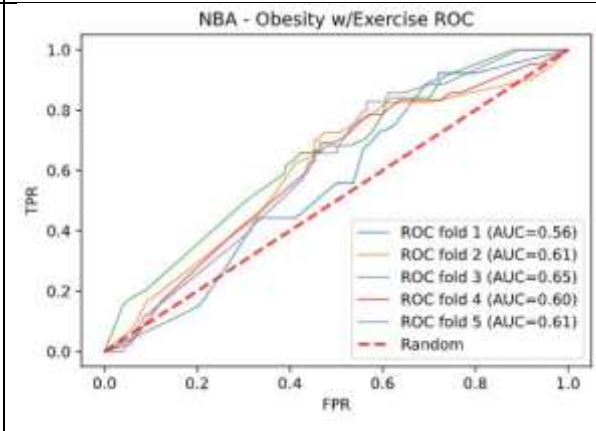
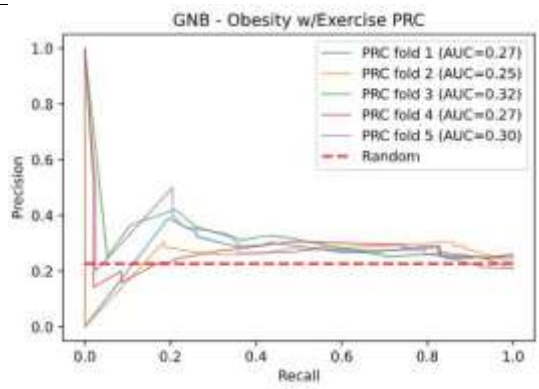
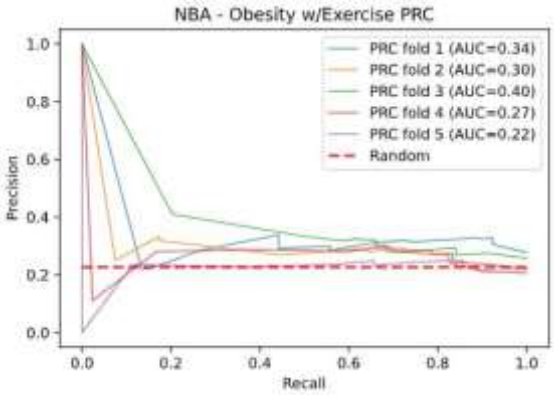
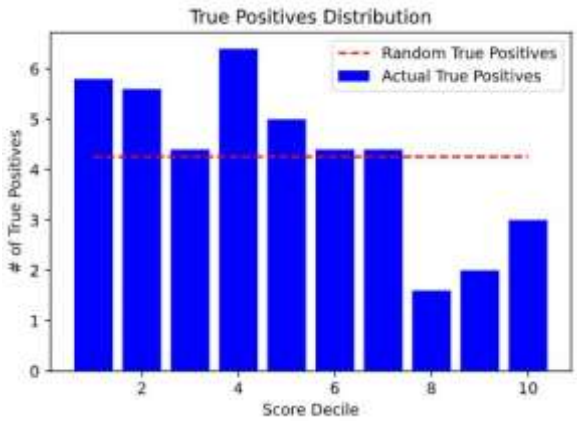
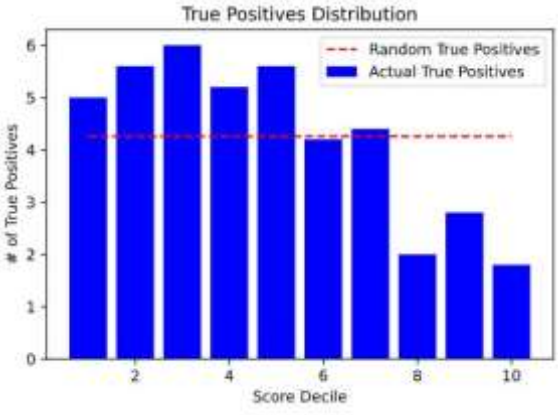
Predicción de *Obesidad* basado en historiales de *Ejercicio*	
GNB	NBA
	
	
	

Tabla. 5.1. Predicción de Obesidad basado en historiales de Ejercicio – Toda la Población

	GNB	NBA
ROC	0.608	0.606
PRC	0.282	0.306

Tabla. 5.2 Promedios de las áreas bajo la curva de las curvas ROC/PRC de las 5 particiones (Obesidad – Ejercicio)

Observando el desempeño de ambos modelos, parece no haber una diferencia clara entre el performance de ambos – si acaso, NBA parece desempeñarse mejor que GNB en términos de precisión - exhaustividad. Esto parece contraponerse a la intuición inicial: GNB no está, a simple vista, desempeñándose mejor que NBA.

Para analizar el porqué de este bajo rendimiento de GNB contra NBA, entendamos primero cómo son ambos modelos diferentes a la hora de clasificar a los individuos de la población. Específicamente, entendamos cómo son diferentes los scores que uno y otro modelo le asignan a una persona. Para hacer esto, tomemos como base a dos individuos p_1 y p_2 cuyas historias de *Ejercicio* en los últimos 20 años son $\boldsymbol{X}^{p_1} = AAAAA$ y $\boldsymbol{X}^{p_2} = BBBBB$, respectivamente. Específicamente, analicemos los scores que GNB y NBA asignarían a cada persona, para distintos rangos de tiempo de referencia:

- Persona $\boldsymbol{p_1}$

Tiempo de referencia	Estructura asociada	Score de GNB	Score de NBA
0 (Sólo información actual)	A****	-0.68933	-0.68933
0 – 1 años	AA***	-1.05614	-1.16307
0 – 5 años	AAA**	-1.03766	-1.34375
0 – 10 años	AAAA*	-1.0164	-1.38656
0 – 20 años	AAAAA	-0.99618	-1.38789

- Persona $\boldsymbol{p_2}$

Tiempo de referencia	Estructura asociada	Score de GNB	Score de NBA
0 (Sólo información actual)	B****	0.335459	0.335459
0 – 1 años	BB***	0.310104	0.615462
0 – 5 años	BBB**	0.277968	0.791717
0 – 10 años	BBBB*	0.166769	0.86838
0 – 20 años	BBBBB	0.179673	0.898681

Los resultados mostrados en las tablas anteriores están presentados gráficamente en la figura siguiente.

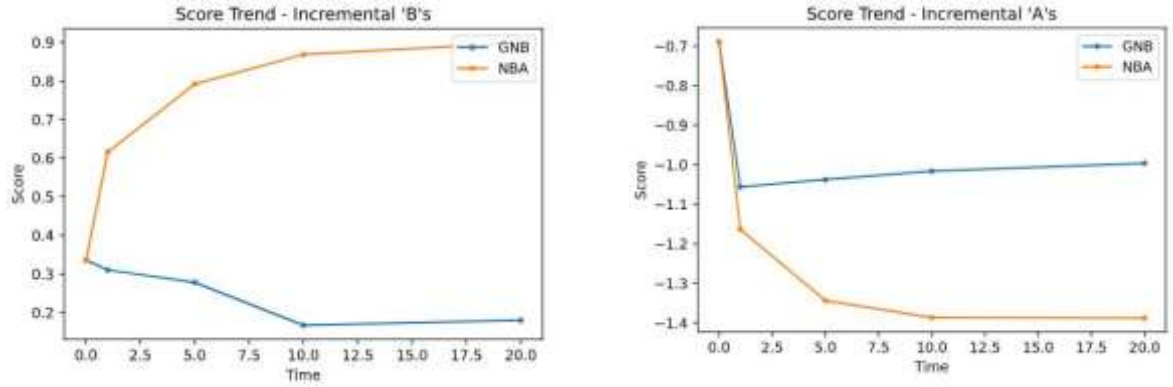


Figura 5.6. Análisis de scores que GNB y NBA asignan a los sub-bloques de información A****, AA***, AAA**, AAAA*, AAAAA, y B****, BB***, BBB**, BBBB* y BBBBB

A partir de esta representación, observamos una diferencia fundamental entre los scores asignados por GNB y NBA. Por un lado, NBA parece *creer* que entre más "B"s hay en la historia de un individuo, más propenso es ese individuo a ser clasificado dentro de *Obesidad* (análogamente, cree que entre más “A”s, más propenso es individuo a ser clasificado dentro de *no-Obesidad*). Es decir, entre más “B”s hay en la historia de un individuo, mayor es su score.

Sin embargo, GNB no parece creer necesariamente lo mismo: entre más “B”s hay en el pasado de un individuo, GNB asigna un score cada vez menor a ese individuo, i.e. para GNB una persona con una historia de “B”s (o “malos” hábitos) sostenidos a lo largo de un tiempo de referencia largo es menos propensa a ser obesa que una persona de la que

sólo se sabe que tiene un hábito malo de ejercicio actualmente. Esto captura el concepto del "*poder* de una historia" de forma satisfactoria: una historia de un hábito "B" sostenido a lo largo del tiempo, dice más de un individuo cuando se analiza como un conjunto, vs. cuando se analiza como una serie de bloques independientes entre sí.

Cuando analizamos este resultado en conjunto con el histograma de distribución de historias en la población general (figura 5.7), encontramos una razón posible del bajo rendimiento de GNB contra NBA cuando los modelos se entrenan usando a la población total: hay muchos patrones de historias "ruidosos", de baja probabilidad de ocurrencia, para los cuales no hay suficiente evidencia como para establecer el patrón/poder de una historia como un conjunto, tal y como sí lo logra hacer en las estructuras AAAAA o BBBBB, de las cuales hay mucha evidencia.

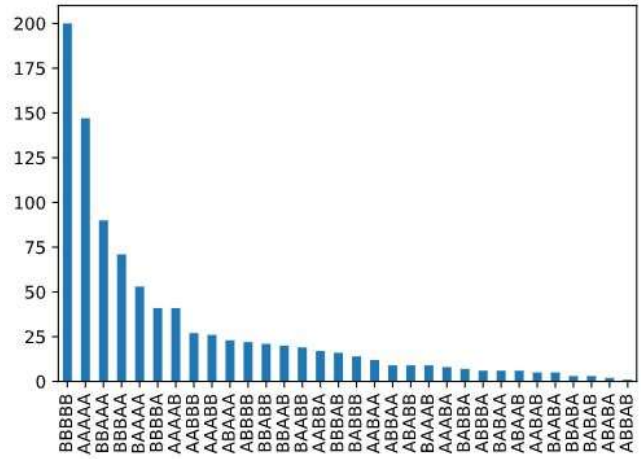


Figura 5.7. Histograma de frecuencias de los patrones de historias de *Ejercicio* en la población de 939 individuos con historias completas de 0 – 20 años.

Por este motivo, además de realizar el proceso de validación cruzada de los modelos usando a la población total de ~930 individuos, lo realizaremos también usando sólo un subgrupo de esta población: aquellos individuos cuyos patrones tienen más de 40 ocurrencias (de manera que nos quedemos con poco más de 2/3 de la población total). Para el caso de los patrones de *Ejercicio*, esto resulta en una población de 643 individuos. Volviendo a hacer la validación cruzada de GNB y NBA sobre este grupo de la subpoblación con historiales de "bajo ruido", obtenemos los siguientes resultados.

Predicción de *Obesidad* basado en historiales de *Ejercicio*	
GNB	NBA

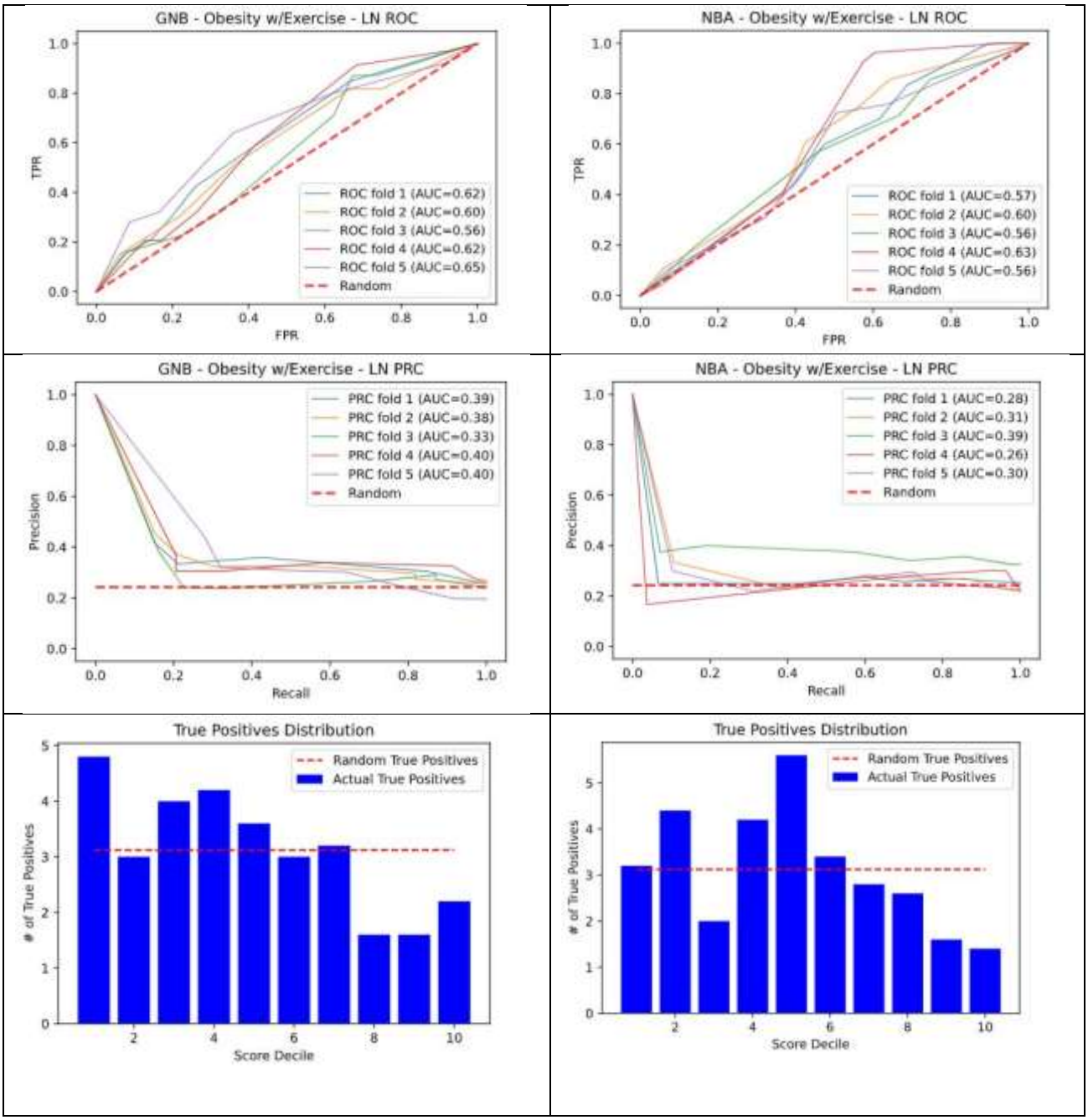


Tabla. 5.3. Predicción de Obesidad basado en historiales de Ejercicio – Población de "historias de bajo ruido"

	GNB	NBA
ROC	0.61	0.584
PRC	0.38	0.308

Tabla. 5.4. Promedios de las áreas bajo la curva de las curvas ROC/PRC de las 5 particiones (Obesidad – Ejercicio)

Observamos que, en esta división de la población, GNB sí es sustancialmente mejor que NBA. Particularmente, GNB es mejor que NBA a la hora de seleccionar verdaderos positivos:

- Por un lado, de los promedios de la tabla 5.4, vemos que GNB muestra una mejora de alrededor de 23.4% en el área bajo la curva de precisión – exhaustividad.
- Además, se observa que la distribución de positivos en GNB es muy efectiva. En el decil de los scores mayores, hay casi un 40% más de positivos de los que se habrían encontrado seleccionando individuos de forma aleatoria.

Estos resultados fortalecen la intuición inicial del poder predictivo de una historia como "conjunto" vs. el análisis de una historia formada de bloques independientes.

5.4 – (Anexo) Otros Resultados

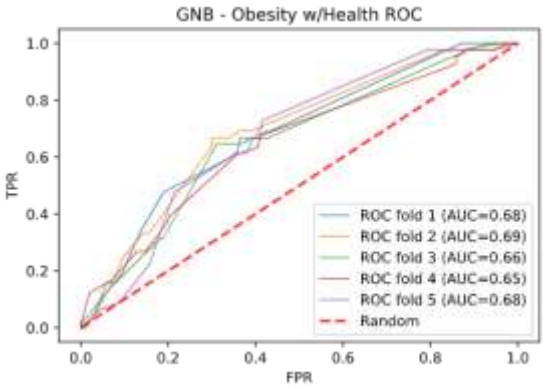
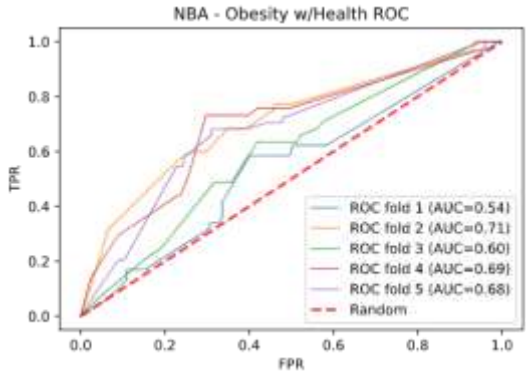
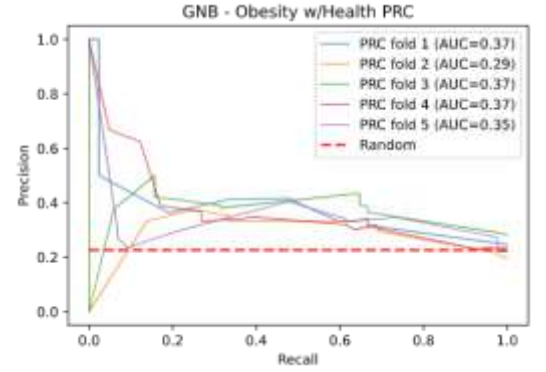
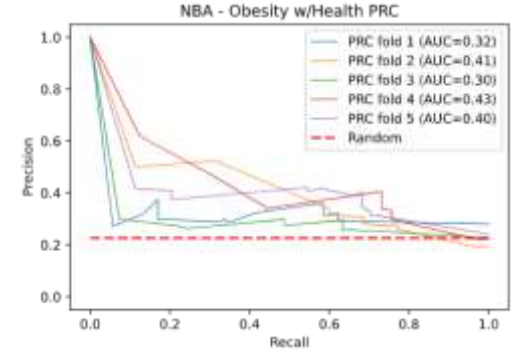
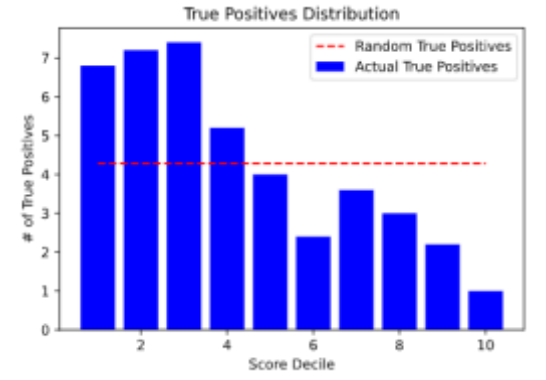
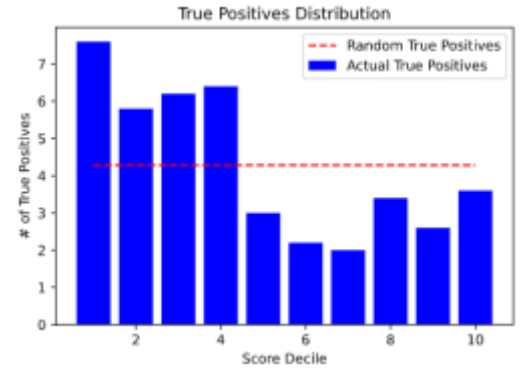
Predicción de *Obesidad* basado en historiales de *Salud*	
GNB	NBA
	
	
	

Tabla. 5.5. Predicción de Obesidad basado en historiales de Salud – Toda la Población

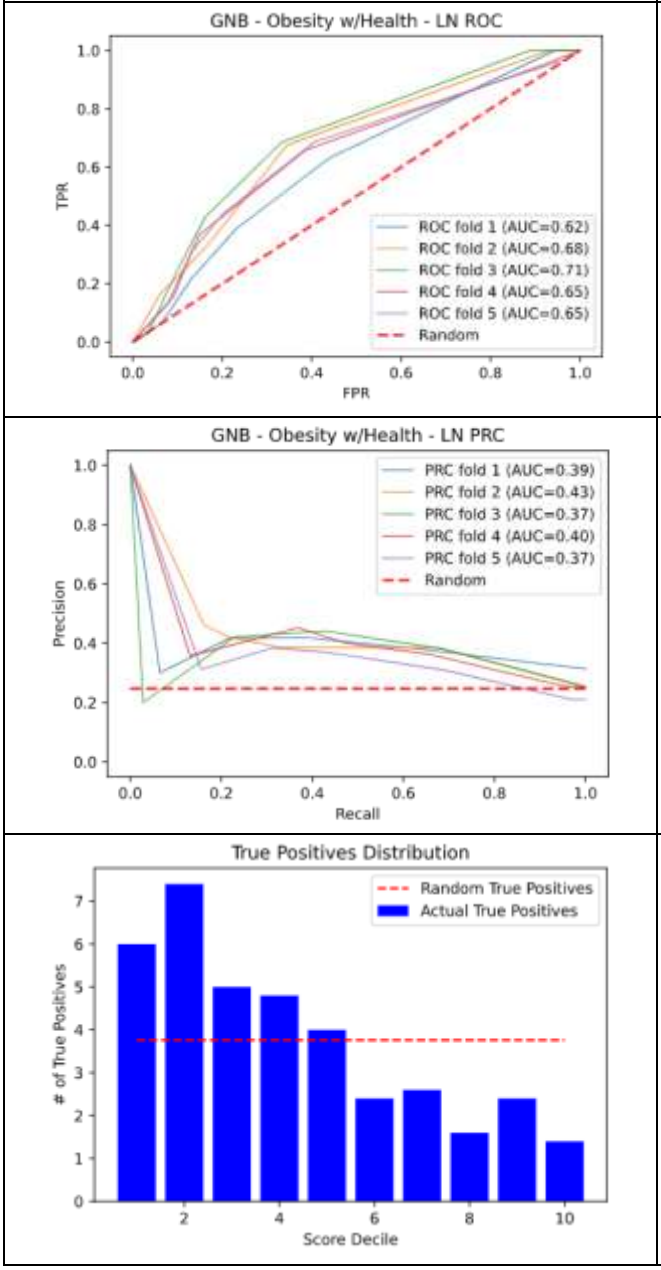
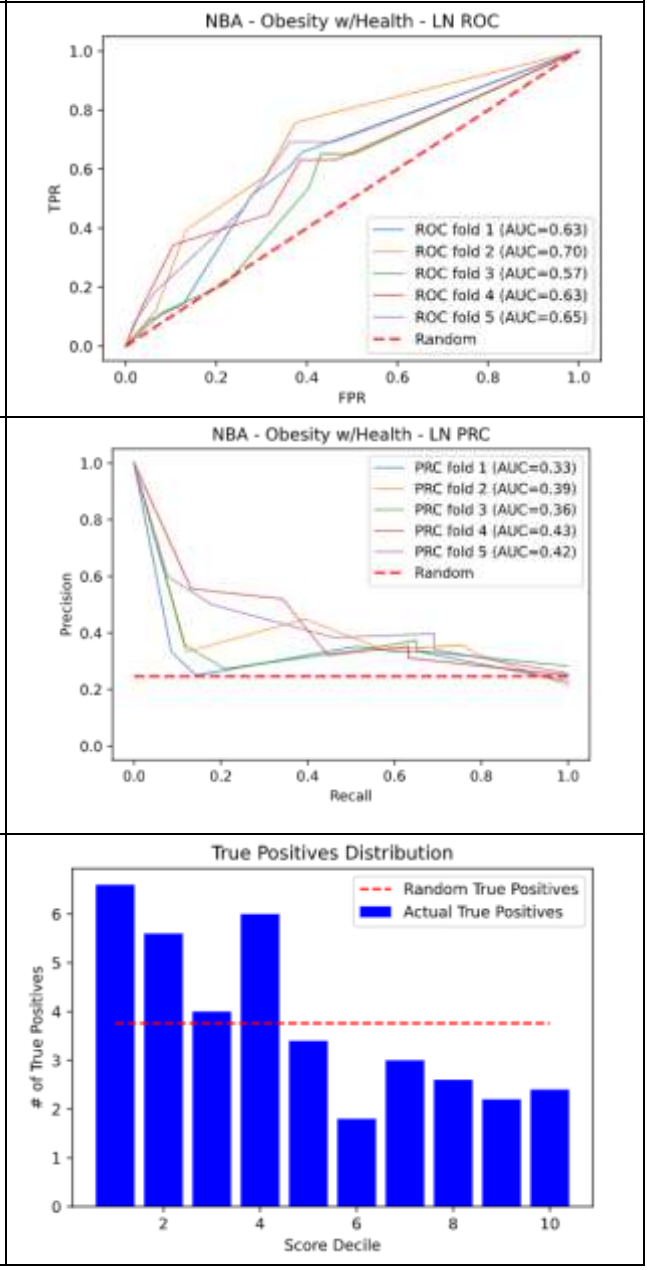
Predicción de *Obesidad* basado en historiales de *Salud*	
GNB	NBA
	

Tabla. 5.6. Predicción de Obesidad basado en historiales de Salud – Población de "historias de bajo ruido"

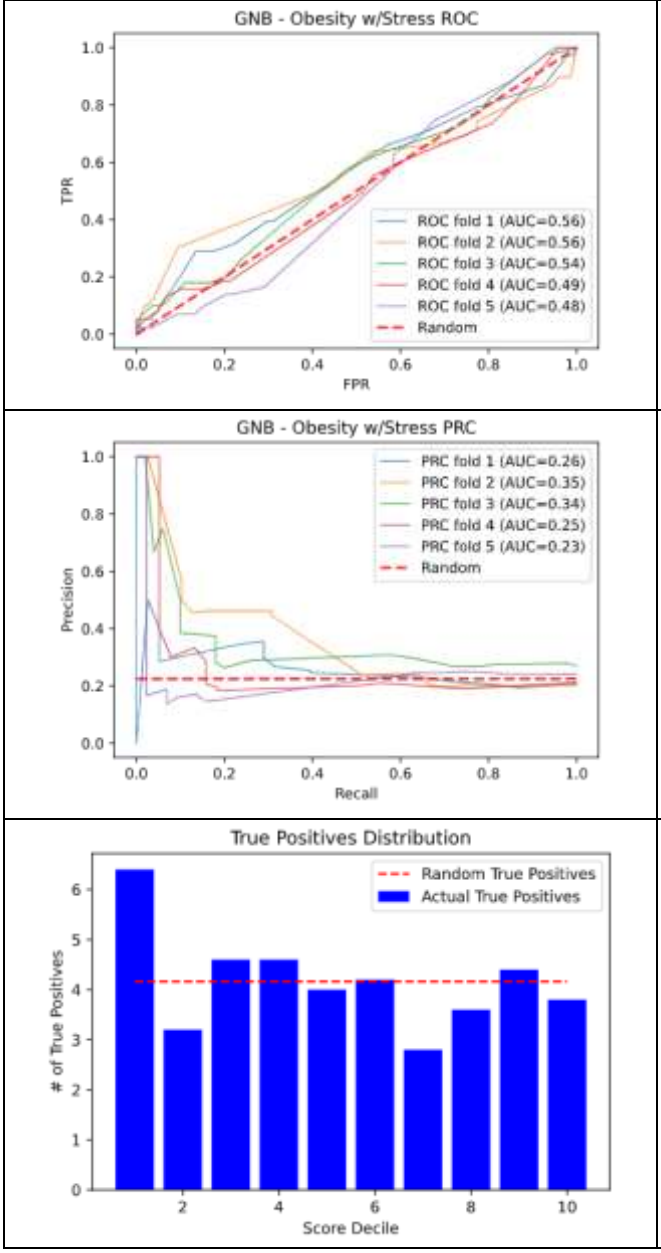
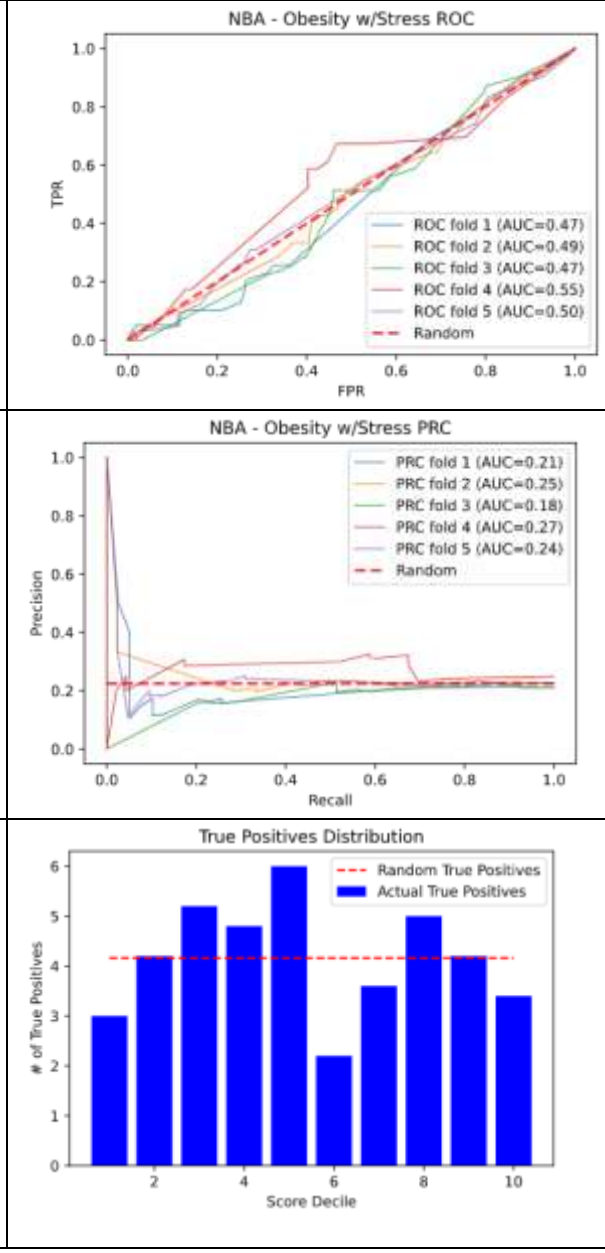
Predicción de *Obesidad* basado en historiales de *Estrés*	
GNB	NBA
	

Tabla. 5.7. Predicción de Obesidad basado en historiales de Estrés – Toda la Población

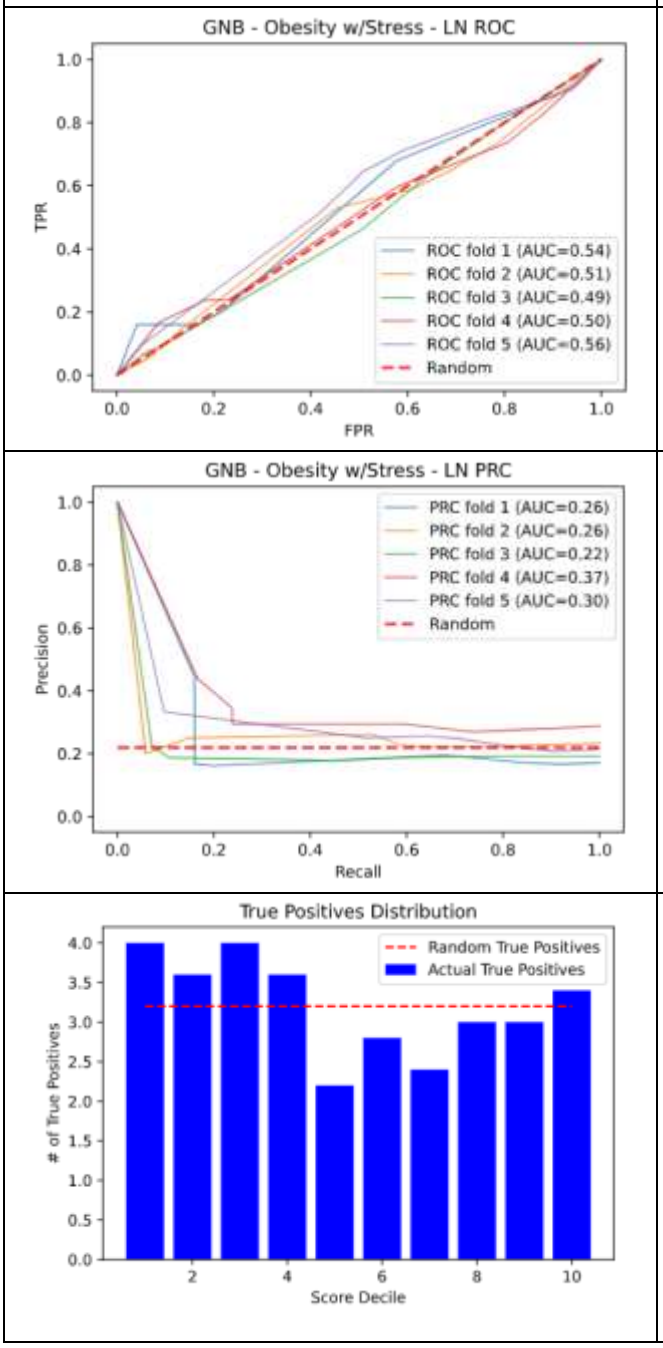
Predicción de *Obesidad* basado en historiales de *Estrés*	
GNB	NBA
	

Tabla. 5.8. Predicción de Obesidad basado en historiales de Estrés – Población de "historias de bajo ruido"

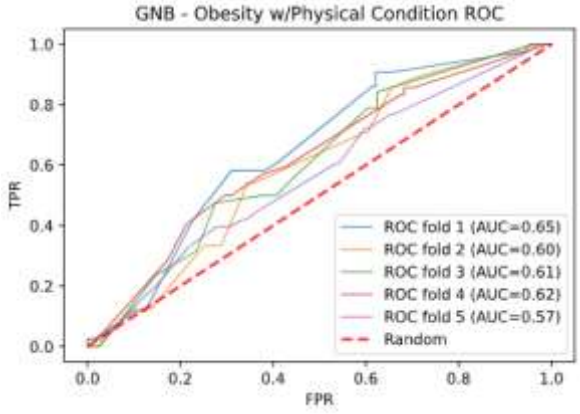
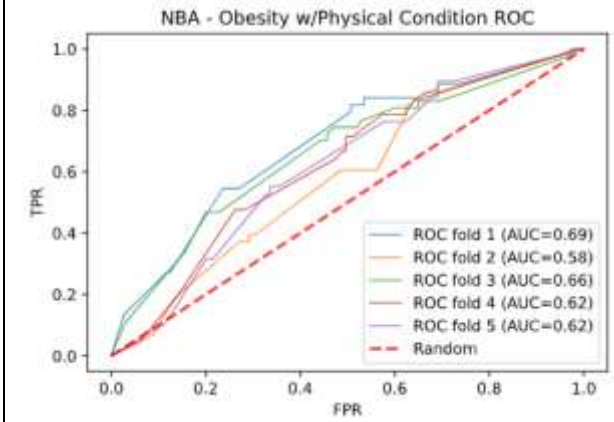
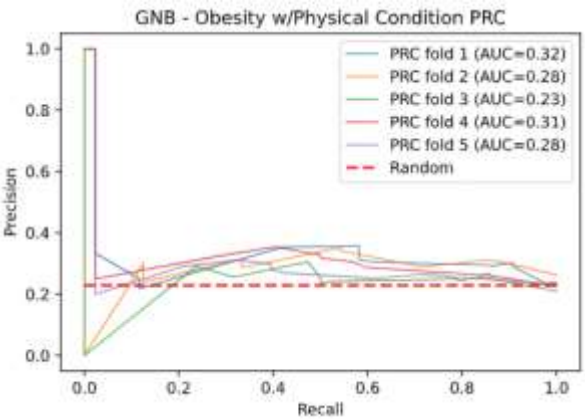
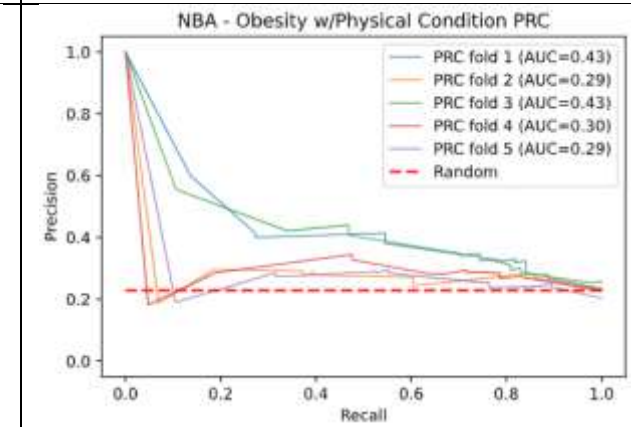
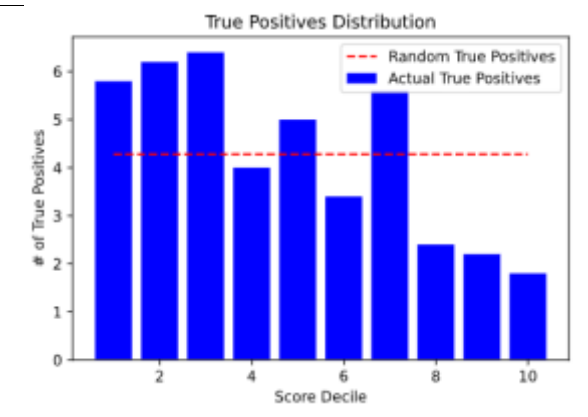
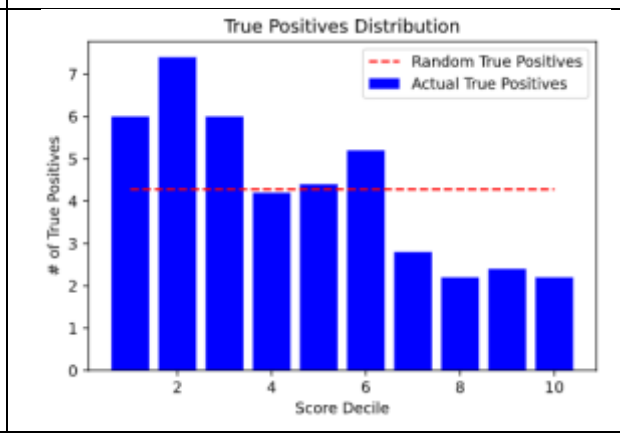
Predicción de *Obesidad* basado en historiales de *Condición Física*	
GNB	NBA
	
	
	

Tabla. 5.9. Predicción de Obesidad basado en historiales de Salud – Toda la Población

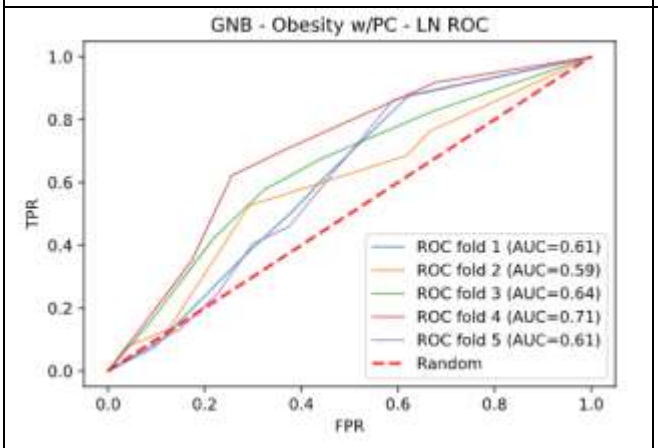
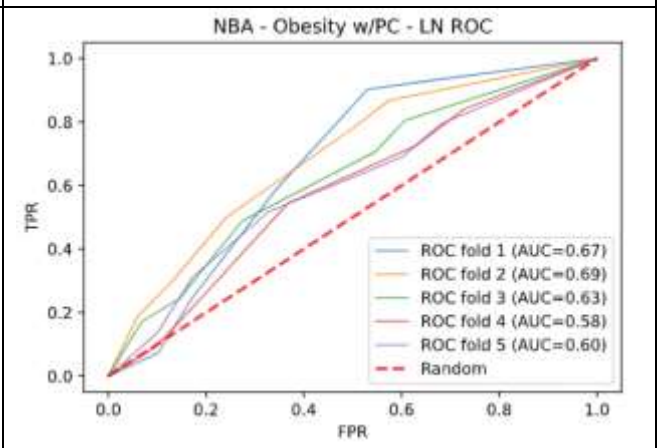
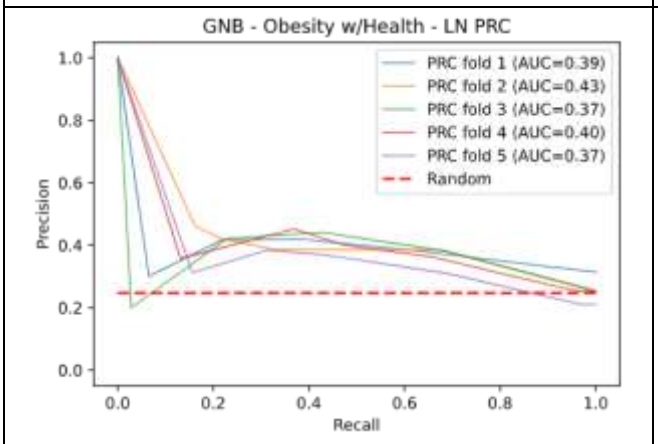
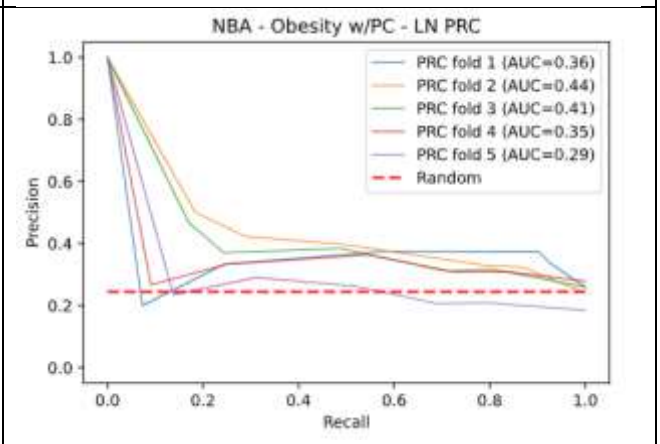
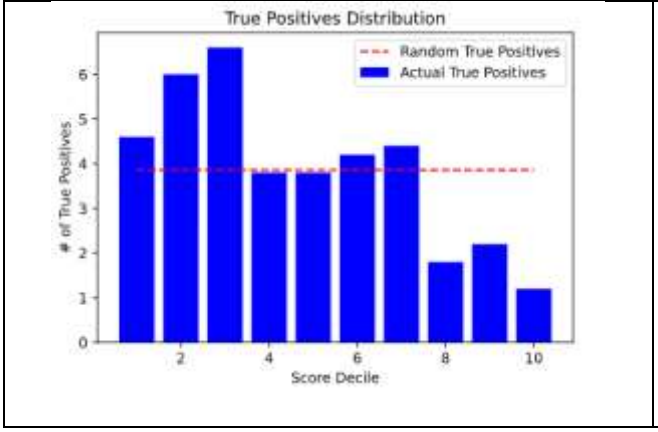
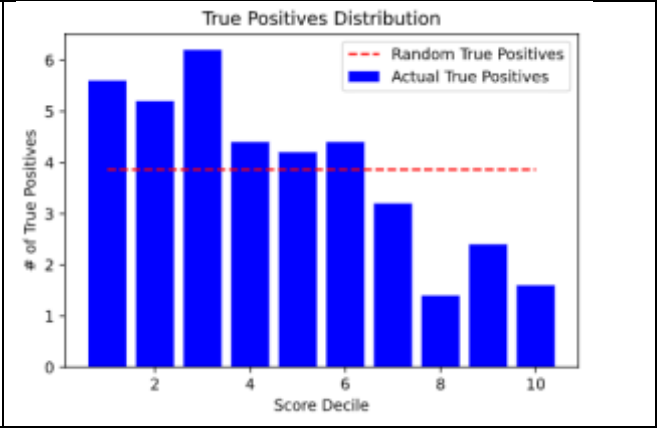
Predicción de *Obesidad* basado en historiales de *Condición Física*	
GNB	NBA
	
	
	

Tabla. 5.10. Predicción de Obesidad basado en historiales de Ejercicio – Población de "historias de bajo ruido"

Capítulo 6. Conclusiones

A continuación, hacemos un sumario de los resultados generales obtenidos a lo largo del trabajo. Los resultados específicos fueron todos discutidos dentro de cada capítulo.

- De los resultados discutidos en el capítulo 3 observamos que, en general, a lo largo de todas las clases analizadas, la fuerza de preservación de hábitos disminuye cuando aumenta el tiempo de referencia, es decir, es más factible preservar hábitos (buenos o malos) en el corto plazo (< 5 años) que en el mediano o largo plazo (> 5 años). Además, observamos que los comportamientos de evolución de hábitos varían cuando se analizan grupos particulares. Por ejemplo, mostramos que:
    - Las personas con obesidad tienden a mantener malos hábitos en mayor medida que las personas sin obesidad, y tienden a mantener buenos hábitos en menor medida que esas últimas.
    - Las personas con alto grado de estudios tienden a mantener buenos hábitos en mayor medida que las personas sin alto grado de estudios, y tienden a mantener malos hábitos en menor medida que estas últimas.
    - Las personas con un puesto académico tienden a mantener buenos hábitos en mayor medida que las personas sin un puesto académico, pero mostraron ser igualmente propensas a mantener malos hábitos que estas últimas.

- De los resultados discutidos en el capítulo 4, mostramos que existen patrones de historias que están muy correlacionados con ciertos grupos. Por ejemplo:
    - Las personas con obesidad están altamente correlacionadas con historias de ejercicio en las que hay un cambio de hábito bueno por uno malo en los últimos 10 años (y luego hay un sostenimiento de un mal hábito). Asimismo, están correlacionadas con una pérdida de una buena condición de salud en los últimos 5 años, y con un historial "inestable" de estrés en los últimos 5 años.
    - Los académicos y las personas con alto grado de estudios están muy correlacionadas con historiales de sostenimiento de condiciones buenas en el mediano plazo (tanto para ejercicio, salud, estrés y condición física), contrario a las personas no-académicas y/o sin alto grado de estudios, que mostraron historiales de manutención de malas condiciones o de cambio hacia malas condiciones en el mismo tiempo de referencia.

- De los resultados discutidos en el capítulo 5, mostramos que:
    - El uso de métodos bayesianos de clasificación cuando hablamos de "historias" como nuestras variables observadas es muy natural e interpretable. En particular, el uso de *GNB* con la factorización utilizada brinda gran intuición al respecto de la evolución del historial de una persona. Al no asumir independencia entre variables, se captura el concepto del "*poder* de una historia" de forma satisfactoria: una historia de un hábito sostenido a lo largo del tiempo dice más de un individuo cuando se analiza como un conjunto, vs. cuando se analiza como una serie de bloques independientes entre sí, y esto se comprobó al analizarse los *scores* asignados por *GNB* y *NBA* a una misma historia.
    - *GNB* mostró ser mejor que *NBA* para encontrar verdaderos positivos dentro de las poblaciones de prueba. Esto fue particularmente cierto para aquellos *scores* muy altos. Estos *scores*, relacionados directamente con las probabilidades que tienen las personas de pertenecer o no a una clase fue particular, son de particular utilidad cuando una de las metas principales de este trabajo es la de generar modelos accionables: si la intención es crear campañas de prevención de obesidad enfocadas, una estrategia viable para hacer esto sería la de enfocar dichas campañas en aquellos grupos cuyos individuos tengan los scores más altos vs. la población general.

Ahora bien, existen varias limitantes con el estudio presentado:

- Quizá la limitante más grande de este estudio es que asume que los resultados reportados para las *Variables de Autoevaluación de Salud* son completamente verídicos. Es decir, asume que todas las personas están respondiendo con veracidad. En una encuesta que hace preguntas sobre el estado de salud de hace 20 o 30 años, no es inapropiado pensar que esta veracidad puede estar condicionada por la memoria o los sesgos de percepción de las personas sobre su pasado.

- Del mismo modo, el hecho de que las cadenas de historias que se usaron para los análisis sean elaboradas con información de hace 20 o 30 años ya presenta un par de problemas en sí:
    - El hecho de que se excluya a la gente más joven de la base de datos.
    - Claramente los historiales de hábitos de los últimos 30 años para un académico de 70 años no son exactamente comparables con el historial de hábitos de un trabajador de vigilancia de 40 años, o de una trabajadora de

administración de 50 años. Las responsabilidades económicas, los impedimentos físicos, el tiempo disponible, etcétera son muy variables dentro de personas de grupos de edad tan distintos. En muchos casos, es probable que sean estos factores los que generan los cambios de hábitos en las personas. En este sentido, el grupo de edad al que una persona pertenece puede ser mejor explicando sus cambios de hábitos que el grado académico, el sexo, el puesto de trabajo o la obesidad.

- La mayor parte de los historiales de las personas no ocurren con suficiente frecuencia como para poder tipificarlas.

- Para la construcción de los modelos de predicción de obesidad, asumimos que la obesidad es producto del historial de hábitos de las personas. No obstante, como se mencionó al inicio de este trabajo, la obesidad es un problema multifactorial. La dieta, la genética y otros factores pueden explicar mejor, en muchos casos, el desarrollo de obesidad.

No obstante, aun con estas limitaciones, debe considerarse que el hacer un estudio longitudinal sobre grupos más homogéneos de edad es de gran dificultad. Si bien un estudio longitudinal es la forma ideal de estudiar cambios en hábitos, debido a esta dificultad, un estudio basado en encuesta como el aquí presentado es más conveniente que no hacer ningún estudio. Estas limitaciones pueden utilizarse como base para abordar este estudio con distintos enfoques.

Con lo anterior en mente, y basados en otras observaciones mencionadas a lo largo de este trabajo, conviene destacar algunos posibles caminos a seguir como trabajo futuro:

- En el capítulo 3, donde encontramos las *fuerzas de preservación y cambio de hábitos*, en general hablamos de estas fuerzas de forma mayormente cualitativa (señalamos tendencias, sin necesariamente cuantificar la magnitud de dichas tendencias). Sin embargo, al observar con detenimiento la evolución de estas fuerzas en diferentes grupos de personas, es posible apreciar que, en diferentes grupos de personas, las tendencias de preservación y cambio tienen diferentes magnitudes. El estudio detallado y cuantificado de estas diferencias entre la magnitud de las tendencias de preservación y cambio de hábitos podría revelar nuevas intuiciones sobre los historiales de hábitos en los distintos grupos de personas estudiados.

- Si bien relacionamos a los historiales de hábitos con los grupos de personas, no buscamos ninguna relación entre historiales de hábitos de distintos tipos.

- Los modelos de predicción de obesidad están basados sólo en un tipo de historial de hábitos (aquellos relacionados con alguno de entre *ejercicio*, *condición física, salud* o *estrés*). Conviene la construcción de un modelo de predicción que tome varias de estas variables (no necesariamente el historial completo) para predecir obesidad.

- Hacer un estudio de hábitos agrupando a la gente por edad y hacer preguntas como:
    - ¿Cómo son diferentes las fuerzas de cambio y preservación de hábitos entre los distintos grupos de edad?
    - ¿A qué edad las personas son más propensas a perder hábitos?

- Mejorar los modelos de *Naive Bayes* presentados en el capítulo 5:
    - Probar con factorizaciones distintas del *likelihood* para los modelos de *Generalized Naive Bayes*. Es decir, crear distintas agrupaciones de variables, de manera que un historial de hábitos pueda ser representado de forma distinta a la presentada en este trabajo.
    - Buscar una forma de lidiar con las historias para las cuales no hay suficiente evidencia como para poder tipificarse (historias "ruidosas"). Probablemente, antes de obtener el score de dicha historia, podría buscarse:
        - La transformación de dicha historia hacia una historia de bajo ruido (transformándola en aquella historia de bajo ruido que más coincidencias tenga).
        - La descomposición de la historia en variables individuales, de forma que su score pueda ser asignado por *NBA* o por alguna otra factorización de *GNB*.

Bibliografía

[1] WHO (2020). *Overweight and Obesity.* https://www.nhlbi.nih.gov/health-topics/overweight-and-obesity

[2] Lee, H., Andrew, M., Gebremariam, A., Lumeng, J. C., & Lee, J. M. (2014). *Longitudinal associations between poverty and obesity from birth through adolescence.* American journal of public health, 104(5), e70–e76.

[3] Rivera, J.A., Colchero, M.A., González de Cosío, T., Aguilar, C., Henández, G., Barquera, S. (2018) *La Obesidad en México.* México: Instituto Nacional de Salud Pública.

[4] Rivera, J.A., Hernández, M., Aguilar, C.A., Vadillo, F., Murayama, C. (2013) *Obesidad en México: Recomendaciones para una Política de Estado.* México: Universidad Nacional Autónoma de México.

[5] Stephens C.R., Gutiérrez J.A.B., Flores H. (2020) *Bayesian Classification of Personal Histories - An application to the Obesity Epidemic.* AMLTA 2019. Advances in Intelligent Systems and Computing, vol 921. Springer, Cham.

[6] Elkan, C. (2001) Magical Thinking in Data Mining: Lessons From CoIL Challenge 2000, 2001. KDD '01: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery.

[7] Yale Stress Center (2016). *The Science of Stress, Bad Habits, and Risk of Chronic Disease.* https://medicine.yale.edu/stresscenter/reduction/science_of_stress_210644_284_25866_v1_372851_284_46027_v1.pdf

[8] Mental Health Foundation (2021). *Stress.* https://www.mentalhealth.org.uk/a-to-z/s/stress

[9] WHO (1948). *Preamble to the Constitution of the World Health Organization as adopted by the International Health Conference.* Geneva: World Health Organization

[10] WHO (2020). *A Healthy Lifestyle.* https://www.euro.who.int/en/health-topics/disease-prevention/nutrition/a-healthy-lifestyle

[11] WHO (2020). *Physical Activity.* https://www.who.int/news-room/fact-sheets/detail/physical-activity

[12] WHO (2020). *Obesity.* https://www.who.int/health-topics/obesity#tab=tab_1

[13] Mind UK (2020). *Causes of Stress*. https://www.mind.org.uk/information-support/types-of-mental-health-problems/stress/causes-of-stress/

[14] Holland, J. (1992). *Adaptation in Natural and Artificial Systems, 2nd edn.* Cambridge: MIT Press.

[15] Stephens CR., Sukumar R. (2006). *An introduction to data mining*. In: Grover R, Vriens M, eds. The Handbook of Marketing Research: Uses, Misuses, and Future Advances, 1st edn. London: Sage Publications.

[16] Easton, J.F., Román Sicilia, H. and Stephens, C.R. (2019). *Classification of diagnostic subcategories for obesity and diabetes based on eating patterns*. Nutr Diet, 76: 104-109.

[17] H. Zhang (2004). *The optimality of Naive Bayes*. Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference. AAAI Press.

[18] C.D. Manning, P. Raghavan and H. Schütze (2008). *Introduction to Information Retrieva*l. Cambridge University Press.

[19] Stephens, C.R., Huerta, H.F. & Linares, A.R. (2018). *When is the Naive Bayes approximation not so naive?* Mach Learn 107, 397–441.

[20] Fawcett T (2006). *An introduction to ROC analysis*. Pattern Recognition Letters, 27(8):861-874.

[21] Sci-Kit Learn Developers (2020). [Gráfico muestra de Precision - Recall] https://scikit-learn.org/stable/_images/sphx_glr_plot_precision_recall_001.png

[22] Sci-Kit Learn Developers (2020). *Precision – Recall*. https://scikit-learn.org/stable/auto_examples/model_selection/plot_precision_recall.html

[23] He, H., Ma, Y. (2013). *Imbalanced Learning: Foundations, Algorithms, and Applications*. Wiley-IEEE Press.

[24] Branco, P., Torgo, L., Ribeiro, R. (2015). *A Survey of Predictive Modelling under Imbalanced Distributions*. Cornell University.

[25] Walber (2014). [Precision – Recall]. https://en.wikipedia.org/wiki/Precision_and_recall#/media/File:Precisionrecall.svg

[26] Sci-Kit Learn Developers (2020) [Gráfico muestra de la curva Receiver Operating Characteristic]. https://scikit-learn.org/stable/_images/sphx_glr_plot_roc_001.png

[27] Ojala, M., Garriga, G. (2010). *Permutation Tests for Studying Classifier Performance*. Journal of Machine Learning Performance.