Universidad Nacional Autónoma de México

Facultad de Ciencias
Licenciatura en Matemáticas Aplicadas
Proyecto 1. Semestre 2023-1

Métodos Estadísticos para la investigación en VIH/SIDA/Salud Pública

Análisis de hábitos de ejercicio para la predicción de obesidad.

Amador González, Noe Eusebio —— 419004815

Profesora Dra. Yanink Neried Caro Vega
Facultad de Ciencias
Universidad Nacional Autónoma de México

Profesora Dra. Lizbeth Naranjo Albarrán
Facultad de Ciencias
Universidad Nacional Autónoma de México

30 de noviembre de 2022
Etapa 3 del proyecto

Índice general

Introducción

1

> *...Paso afirmando, paso negando, paso con dudas*
> *Entre risas y amarguras, buscando el por qué y*
> *el cuándo...*
>
> **— Rubén Blades**

1.1. Motivación y Pregunta de Investigación

La obesidad y el sobrepeso afectan al 39 % de los niños, 44 % de los adolescentes y 74 % de las personas adultas en el país [7]. Se sabe que existe una alta asociación entre la obesidad y múltiples enfermedades, como diabetes tipo 2, cáncer de mama y de colon, problemas respiratorios, presión arterial alta, colesterol total y triglicéridos altos, entre otros [@9].

El proyecto Conductome es una consecuencia del interés por estudiar los motivos y causas del desarrollo de sobrepeso y obesidad en la población de México. Busca respuestas y posibles soluciones tomando en cuenta la mayor cantidad de puntos de vista por parte de diferentes áreas de estudio que contribuyan a un análisis más completo y ambicioso. Conductome considera que la obesidad y el sobrepeso son problemas complejos que requieren análisis complejos para entenderlos, pues dichos problemas no parecen ser derivados de una única razón, sino que es un conjunto de conductas, decisiones e interacciones entre organismos las que guían al desarrollo de estos problemas. [@2, @4, 3]

En este sentido, una de las preguntas que nos hacemos al intentar entender este problema es:

¿Podemos predecir obesidad dado nuestro historial de hábitos de ejercicio?

1.2. Objetivos

1.2.1. Objetivos Generales (del proyecto Conductome)

Los problemas y enfermedades derivados del sobrepeso y la obesidad nos afectan desde lo individual hasta lo social. Es por esto que debemos establecer nuevos planes y metas que puedan ayudarnos y motivarnos a combatir estas condiciones.

Algunos de los objetivos establecidos en el proyecto Conductome son:

- Identificar y analizar las relaciones causales entre los factores que influyen en la conducta humana relacionada con la obesidad.
- Identificar las razones detrás de las conductas nocivas relacionadas con la obesidad y construir un modelo mental[1], que tome en cuenta los conceptos como creencias, percepción de sí mismo y del mundo.
- Proponer políticas públicas y estrategias para incidir en la reducción del desarrollo de enfermedades del síndrome metabólico.

1.2.2. Objetivos Específicos (de mi proyecto)

Los objetivos establecidos para el desarrollo de este proyecto son:

- Analizar, cuestionar y replicar o mejorar los resultados mostrados en [1] para responder a nuestra pregunta de investigación.
- Aplicar un modelo distinto al utilizado en [1] para comparar ambos modelos y buscar diferencias y similitudes.
- Presentar resultados/respuestas a la pregunta de investigación y justificar qué modelo funcionó mejor.

[1] Un modelo mental es un mecanismo del pensamiento mediante el cual un ser humano intenta explicar cómo funciona el mundo real.

1.3. Antecedentes

El proyecto *Conductome* inició sus estudios en 2014, cuando se recabó información de aproximadamente 4,000 personas pertenecientes a la comunidad UNAM, tanto estudiantes como trabajadores académicos y no académicos.

Actualmente el proyecto cuenta con una base de datos que permite planear estudios longitudinales para estudiar la evolución de los participantes ya que, afortunadamente, el proyecto se ha mantenido activo durante estos años de manera que en 2019 y en 2022 se realizaron actualizaciones de la información de la mayoría de las personas a las que se les realizaron los estudios y cuestionarios en 2014.

Algunos estudios realizados a los participantes tuvieron como objetivos conocer los niveles de glucosa, trisgliceridos, colesterol, ácido úrico, niveles séricos de creatinina, entre otros.

Para objetivos de este proyecto nos interesa estudiar los datos relacionados a *hábitos de ejercicio* que las personas brindaron por medio del cuestionario en la sección *Estilo de vida*. La base de datos, las secciones y las variables las describiré en el siguiente capítulo.

Durante estos años se han encontrado resultados muy interesantes que pueden ser consultados en la página web del proyecto: *chilam.c3.unam.mx*

Sin embargo, con el objetivo de resaltar algunos resultados que esperamos respaldar en el trabajo que desarrollaremos en esta materia podemos mencionar:

- 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 estas últimas [1].
- 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 [1].

En este proyecto esperamos coincidir con algunos de estos resultados y respaldarlos con nuestro análisis.

1.4. Justificación

Una de las justificaciones a este trabajo es notar el impacto que este problema tiene en el mundo, pero principalmente en nuestro país, pues se ha mostrado que alrededor del 72.5 % de los adultos del país tiene sobrepeso. [5].

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.[5].

La pérdida económica -en términos de productividad- del país debido a esta condición se estima en 71 669 millones de pesos (0.4 % del PIB) por año. [5].

Y una de las más importantes razones para seguir estudiando y buscando solución a estas condiciones es que se estima que aproximadamente el 25.3 % del total de las muertes en el país en 2004 fueron consecuencia de problemas relacionados a sobrepeso y obesidad [6].

Es por esto que considero relevante y útil difundir y respaldar trabajos que incentiven interés en este tema.

Planteamiento del Problema

2

> *I'll never learn if I never leap*
> *I'll always yearn if I never speak*
>
> — **Adele**

2.1. Análisis Descriptivo

2.1.1. Base de Datos

La base de datos fue construida por medio de cuestionarios y estudios de laboratorio realizados a 1,073 participantes entre alumnos, personal académico y personal no académico de la UNAM bajo el proyecto FM/DI/023/2014 usando protocolos aprobados por el Comité de Ética de la Facultad de Medicina de la UNAM.

Las variables que vamos a utilizar en este proyecto corresponden a variables de *autoevaluación de la salud*, *estilo de vida* y *antropometría*.

- **Datos personales**
  - *id_sexo*: Correspondiente al sexo del participante.
    {*Hombre, Mujer)*}
  - *Aedad*: Edad del participante
  - *AAedad*: Grupo de edad del participante
    {*19-27, 28-32, 33-37, 38-42, 43-47, 48-52, 53-58, 59-81*}
  - *id_gestud*: Grado de estudios
    {*Prim, Sec, Bach, CarTec, Lic, Mast, Doc, PDoc, Otro*}
- **Estilo de vida**
  - *ejer_act*, *ejer1*, *ejer5*, *ejer10*: Número de horas de ejercicio por semana actualmente, hace 1 año, hace 5 años y hace 10 años.

- **Autoevaluación de la salud**

  - *salud_act*, *salud1*, *salud5*, *salud10*: Percepción de la salud actualmente, hace 1 año, hace 5 años y hace 10 años. [1-8]

  - *peso_act*, *peso1*, *peso5*, *peso10*: Percepción del peso actualmente, hace 1 año, hace 5 años y hace 10 años. [1-8]

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

Tab. 2.1: Tabla de valores para las variables de percepción de salud y peso.

  - *peso*: Peso actual del participante en kilogramos.

  - *estatura*: Estatura actual del participante en centimetros.

  - *pess_acc*: Acciones a tomar por el peso.[1-5]

Valor	Significado
1	Bajar de peso
2	Estás contento
3	Subir de peso
4	No sé
5	No quiero responder

Tab. 2.2: Tabla de valores para la variable *pess_acc*

- **Antropometría**

  - *IMC*: Índice de Masa Corporal.

  - *AIMC*: Categoría de IMC [1-6].

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

Tab. 2.3: Tabla de valores para la variable *AIMC*.

2.1.2. Análisis Exploratorio de la Base de Datos

En la figura 2.1 podemos ver que la mayoria de los participantes son mujeres. También notamos que los *grupos de edades* son similares y que la mayoría de los participantes tienen un *grado de estudios* de licenciatura.

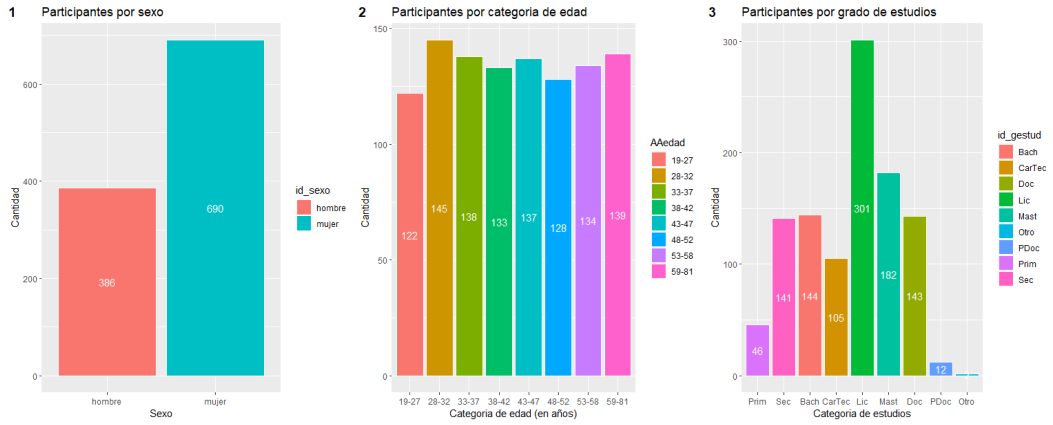
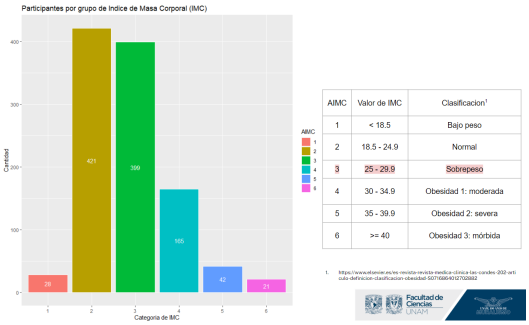


Fig. 2.1: Participantes por *sexo*, *grupo de edad* y *grado de estudios*.

En la figura 2.2 podemos observar que la mayoría de los participantes están en el grupo 2 de *AIMC* y esto significa que tienen un peso normal, sin embargo, considerando a los participantes clasificados en los grupos 3, 4, 5 y 6 podemos notar que todos ellos pertenecen a los grupos de interés para este trabajo porque son aquellas personas que tiene sobrepeso u obesidad. En pocas palabras, podemos ver que la mayoría de los participantes tienen sobrepeso u obesidad.

En la figura 2.3 podemos observar que los grupos de edad con mayores porcentajes de sobrepeso u obesidad son los que están en el rango de 38 a 81 años, mientras que los más jóvenes presentan porcentajes mayores de peso normal. También podemos ver que en todos los grupos la mayoría de las personas quiere bajar de peso. Aunque es interesante notar que las personas del grupo [59-81] aumenta el porcentaje de satisfacción con su peso y disminuye su deseo de bajar de peso.



AIMC	Valor de IMC	Clasificacion[1]
1	< 18.5	Bajo peso
2	18.5 - 24.9	Normal
3	25 - 29.9	Sobrepeso
4	30 - 34.9	Obesidad 1: moderada
5	35 - 39.9	Obesidad 2: severa
6	>= 40	Obesidad 3: mórbida

1. https://www.elsevier.es/es-revista-revista-medica-clinica-las-condes-202-articulo-definicion-clasificacion-obesidad-S0716864016702882

Fig. 2.2: Cantidad de participantes en cada grupo de AIMC.

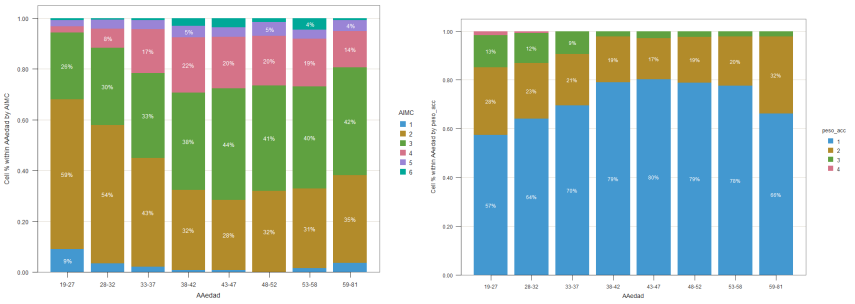


Fig. 2.3: Porcentajes de categoría AIMC divididos por grupos de edad (izquierda). Acciones a tomar por grupo de edad (derecha).

En la figura 2.4 podemos ver que la gran mayoría de las personas hace 0 horas de ejercicio a la semana, y el resto se distribuye entre 1 y 5 horas a la semana. También observamos que hace 10 y 5 años solía haber más personas que hacían 5 horas de ejercicio e incluso bastantes reportaron hacer hasta 10 horas de ejercicio a la semana en el pasado, sin embargo, en la actualidad la cantidad de personas que reportaron 10 horas de ejercicio a la semana es mucho menor.

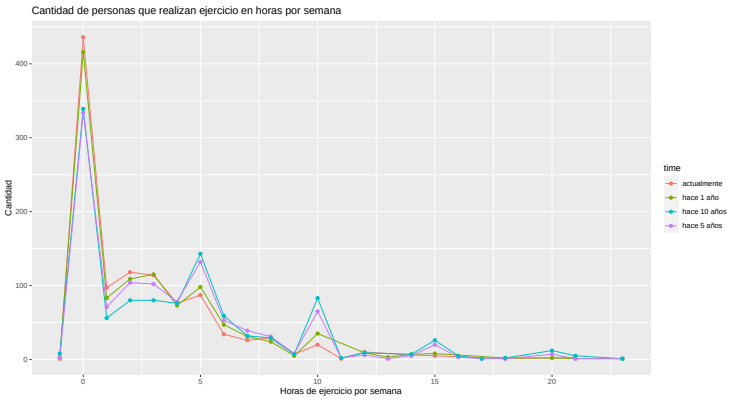


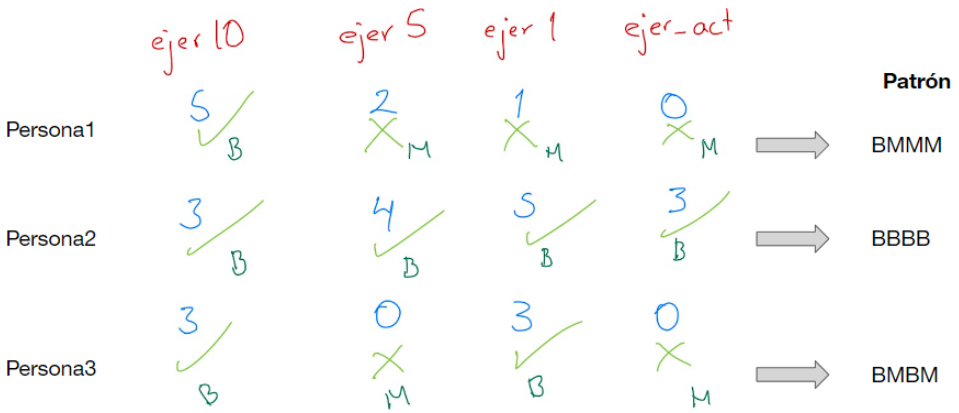
Fig. 2.4: Horas de ejercicio actual, hace 1, 5, y 10 años.

2.2. Identificación del Modelo Estadístico

Con todo lo anterior en mente, debemos pensar cómo relacionar el historial de hábitos de ejercicio con predecir si una persona puede tener sobrepeso u obesidad.

En primer lugar recordemos que la Organización Mundial de la Salud recomienda realizar 2.5 horas de ejercicio a la semana como mínimo [@10], esto significará para este estudio que si una persona realiza menos de 2.5 horas de ejercicio a la semana entonces tiene un *mal hábito de ejercicio (M)*, en caso contrario tendría un *buen hábito de ejercicio (B)*.

Si tenemos 6 variables que describen el tiempo de ejercicio que las personas realizaban en diferentes tiempos, podemos ver algo así:



Donde, por ejemplo, la ***Persona3*** con el patrón ***BMBM*** indica que hace 10 años tuvo un buen hábito de ejercicio, hace 5 años tuvo un mal hábito de ejercicio, hace 1 año tuvo un buen hábito de ejercicio y actualmente tiene un mal hábito de ejercicio.

De esta manera se generar un patrón de ejercicio correspondiente para cada persona en nuestra base de datos.

Y dado que sabemos cuál es la situación actual de dicha persona, es decir, si tiene obesidad o no, entonces para resolver este problema se podría considerar un modelo/algoritmo de clasificación binaria que nos permite predecir el estado de una persona (obesidad=Si, obesidad=No) dado su patrón de ejercicio en los últimos 10 años.

En [1] utilizan un clasificador bayesiano ingenuo (Naive Bayes) para lograr esta clasificación. **Yo propongo utilizar una Regresión Logística y comparar los resultados de ambos modelos.**

Método y Resultados

3

> *There's more to be seen than can ever be seen, more to do than can ever be done.*
>
> **— Tim Rice**

3.1. Preprocesamiento de datos

Considerando la base de datos descrita en 2.1.1 tuvimos que realizar un procesamiento de las variables *ejer_act*, *ejer1*, *ejer5* y *ejer10* para obtener el patrón descrito en 2.2 correspondiente para cada participante.

De las misma manera se creó la variable indicadora de *obesidad* para cada persona tomando en cuenta su índice de masa corporal capturado.

Cabe resaltar que se realizó un filtrado para que las variables *ejer_act*, *ejer1*, *ejer5* y *ejer10* no fueran nulas, por lo que pasamos de tener 1,073 registros a tener 1,067.

La base de datos con 1,067 registros a utilizar para entrenar nuestros modelos tiene la siguiente estructura:

dp_folio	id_sexo	Aedad	id_gestud	ejer_0	ejer_1	ejer_5	ejer_10	ejer0_10	obesity
1	mujer	51	CarTec	M	M	M	M	MMMM	1
2	mujer	38	Bach	M	M	B	B	MMBB	0
3	mujer	34	Sec	M	M	M	B	MMMB	1
4	hombre	63	CarTec	M	M	B	B	MMBB	1
5	hombre	42	Sec	M	M	M	M	MMMM	0

Tab. 3.1: Ejemplo de datos preprocesados para utilizar en los modelos.

Donde la variable *obesity* es la variable que queremos predecir.

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 $X = (x_1, x_2, x_3, x_4)$ un vector de observaciones tales que x_i corresponde al valor de la variable de hábito de ejercicio a_t para algún tiempo t (donde $a_t \in \{B, M\}$ y $t \in \{0, 1, 5, 10\}$).

Por ejemplo: $X = (x_1 = a_0, x_2 = a_1, x_3 = a_5, x_4 = a_{10}) = (B, M, M, M)$ corresponde a una persona que actualmente tiene buenos hábitos de ejercicio pero hace 1, 5 y 10 años no los tenía.

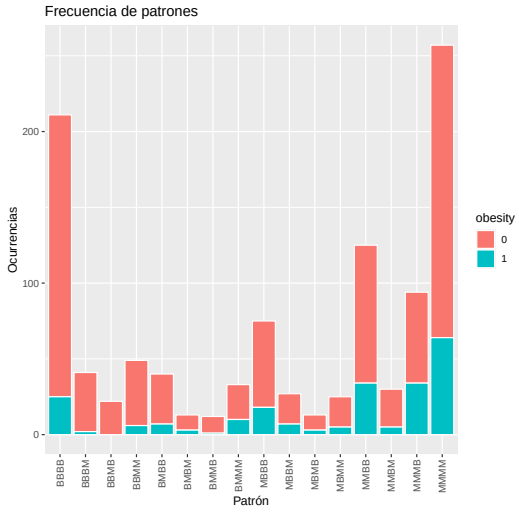


Fig. 3.1: Frecuencia de patrones de hábitos de ejercicio coloreado por clase obesidad y no obesidad.

En la figura 3.1 observamos que, con los datos ya procesados, los patrones más comunes son *MMMM* y *BBBB*. También se puede puede observar que las personas que pertenecen a la clase obesidad suelen tener un *mal hábito de ejercicio* actualmente.

3.2. Modelo Naive Bayes

Este modelo se utiliza con el motivo de presevar la idea mostrada en [8].

Naive Bayes está basado en el *Teorema de Bayes* y asume independencia entre las variables explicativas. Por lo que tomaremos en cuenta que:

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)} \tag{3.1}$$

De donde podemos evitar calcular $P(X)$ considerando el siguiente cociente y $\bar{C}$ la clase complemento de C:

$$\frac{P(C|X)}{P(\bar{C}|X)} = \frac{\frac{P(X|C)P(C)}{P(X)}}{\frac{P(X|\bar{C})P(\bar{C})}{P(X)}} = \frac{P(X|C)P(C)}{P(X|\bar{C})P(\bar{C})} \tag{3.2}$$

De manera que Stephens et. al define la función score en [8] como:

$$S(C|X) = \ln\left(\frac{P(C|X)}{P(\bar{C}|X)}\right) = \sum_{i=1}^{n} S(X_i) + \ln\left(\frac{P(C)}{P(\bar{C})}\right) \tag{3.3}$$

Donde

$$S(X_i) = \ln\left(\frac{P(X_i|C)}{P(X_i|\bar{C})}\right) = \ln\left(\frac{\frac{N_{CX_i}}{N_C}}{\frac{N_{X_i}-N_{CX_i}}{N-N_C}}\right) \approx \ln\left(\frac{\frac{N_{CX_i}+1}{N_C+2}}{\frac{N_{X_i}-N_{CX_i}+1}{N-N_C+2}}\right) \tag{3.4}$$

Con esta ecuación, clasificamos a cada una de las personas dentro de la clase C si $S(C|X) > 0$ o clasificamos dentro de la clase $\bar{C}$ si $S(C|X) < 0$.

N es el número total de elementos en la muestra. N_C es el número de elementos que pertenecen a la clase obesidad. N_{X_i} es el número total de elementos que cumple la característica $X = x$ en la muestra. N_{CX_i} es el número de elementos que cumple la condición $X = x$ y que pertenecen a la clase obesidad.

Por ejemplo: Tomando el historial mostrado anteriormente $X = (B, M, M, M)$ el score asociado sería

$$S(C=1|X) = \sum_{i=1}^{4} S(X_i) = S(a_0 = B) + S(a_1 = M) + S(a_5 = M) + S(a_{10} = M) + \ln\left(\frac{P(C)}{P(\bar{C})}\right)$$

$$= \ln\left(\frac{\frac{170+1}{224+2}}{\frac{421-170+1}{1067-224+2}}\right) + \ln\left(\frac{\frac{158+1}{224+2}}{\frac{604-158+1}{1067-224+2}}\right) + \ln\left(\frac{\frac{123+1}{224+2}}{\frac{505-123+1}{1067-224+2}}\right) + \ln\left(\frac{\frac{102+1}{224+2}}{\frac{475-102+1}{1067-224+2}}\right) + \ln\left(\frac{0.2099}{0.79}\right)$$

$$= \ln(2.5371) + \ln(1.3299) + \ln(1.2105) + \ln(1.0297) + \ln(0.2656)$$

$$= 0.9310 + 0.2851 + 0.1910 + 0.0292 - 1.3257 = 0.1105 > 0$$

Por lo que, según este modelo, una persona con patrón *BMMM* se clasificaría dentro de la clase obesidad.

3.3. Modelo de Regresión Logística

La *regresión logística* también es un método muy utilizado para clasificación binaria. Se utiliza la función liga *Logit* (η) y una distribución binomial como componente aleatorio y las variables *ejer_act*, *ejer1*, *ejer5* y *ejer10* como componente sistemático.

De forma general se ve como:

$$\eta = X\beta \tag{3.5}$$

Para estimar la probabilidad de pertenecer a la clase obesidad se tomaron en cuenta dos modelos:

El primer modelo corresponde a un modelo con interacciones entre las variables predictoras:

$$\eta = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 ejer\_act + \beta_2 ejer1 + ... + \beta_{15} ejer\_act * ejer1 * ejer5 * ejer10 \tag{3.6}$$

donde $p = P(C = 1|X = x)$. Este modelo estima 16 parámetros/coeficientes:

```
Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                        -2.2437     0.2811  -7.983 1.43e-15 ***
ejer_0M                             1.3863     0.4251   3.261  0.00111 **
ejer_1M                             0.8575     0.5736   1.495  0.13494
ejer_5M                           -15.3223  1318.7268  -0.012  0.99073
ejer_10M                          -15.3223   791.2361  -0.019  0.98455
ejer_0M:ejer_1M                    -0.9754     0.7004  -1.393  0.16376
ejer_0M:ejer_5M                    14.5703  1318.7273   0.011  0.99118
ejer_1M:ejer_5M                    14.6292  1318.7273   0.011  0.99115
ejer_0M:ejer_10M                   15.1682   791.2364   0.019  0.98471
ejer_1M:ejer_10M                   15.0992   791.2370   0.019  0.98477
ejer_5M:ejer_10M                   30.5858  1537.8866   0.020  0.98413
ejer_0M:ejer_1M:ejer_5M           -13.6185  1318.7279  -0.010  0.99176
ejer_0M:ejer_1M:ejer_10M          -15.4047   791.2375  -0.019  0.98447
ejer_0M:ejer_5M:ejer_10M          -29.9208  1537.8873  -0.019  0.98448
ejer_1M:ejer_5M:ejer_10M          -28.7352  1537.8875  -0.019  0.98509
ejer_0M:ejer_1M:ejer_5M:ejer_10M   28.3303  1537.8883   0.018  0.98530
```

El segundo modelo corresponde a un model sin interacciones en las variables predictoras:

$$\eta = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 ejer\_act + \beta_2 ejer1 + \beta_3 ejer5 + \beta_4 ejer10 = X\beta \tag{3.7}$$

donde $p = P(C = 1|X = x)$. Este modelo estima 5 parámetros/coeficientes:

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.1466     0.2084 -10.299  < 2e-16 ***
ejer_0M       0.8466     0.2598   3.258  0.00112 **
ejer_1M       0.4089     0.2532   1.615  0.10626
ejer_5M       0.1946     0.2350   0.828  0.40773
ejer_10M     -0.1727     0.2197  -0.786  0.43193
```

En ambos casos observamos un comportamiento similar, sin embargo el segundo modelo arrojó resultados ligeramente mejores como veremos en la sección 3.4.

Es importantes destacar que según los dos modelos ajustados, se pudo observar que dada la variable de ejercicio actual (*ejer_0*) las demás variables parecen no ser tan significativas. Esto nos estaría diciendo un poco que la variable *ejer_0* de ejercicio actual sería muy influyente y clave para predecir el estado en que la persona se encuentre, es decir si tiene obesidad o no.

También probé un modelo con todas las variables del dataframe mostrado en la tabla 3.1 sin interacciones. De aquí pude notar que la variable *id_gestud* era significante y se podría plantear la hipótesis de que incluir esta variable en el modelo podría mejorar su rendimiento. Es decir, que el grado de estudios también parece ser un factor importante para predecir obesidad. Desafortunadamente ya no pude desarrollar más este análisis.

En nuestro caso, asignaremos una observación a la clase obesidad si $P(C = 1|X) > 0.5$ y a la clase no obesidad en otro caso. Queda pendiente estudiar si existe un punto de corte más adecuado que permita clasificar con mejor precisión.

3.4. Comparación de Modelos y Resultados

Una vez que se entrenaron los modelos, los resultados fueron los siguientes:

	Naive Bayes	RegLogistica con interacciones	Reg Logistica sin interacciones
precisión	85 %	81 %	81 %
AUC	0.61	0.579	0.582

Tab. 3.2: Tabla de precisiones y AUCs.

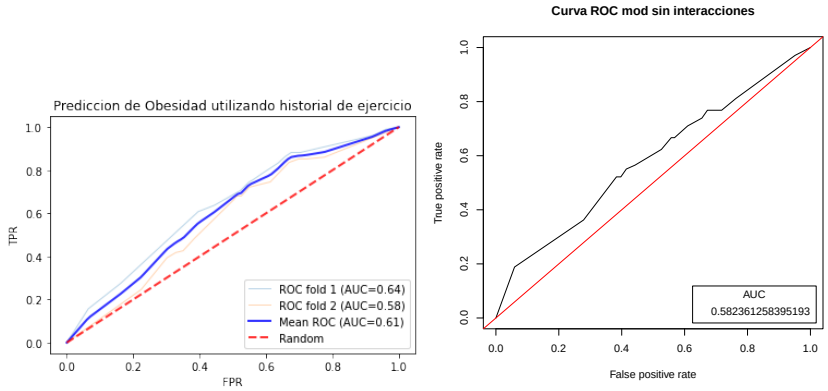


Fig. 3.2: Curvas ROC para el clasificador Naive Bayes y para el modelo de regresión logística sin interacciones.

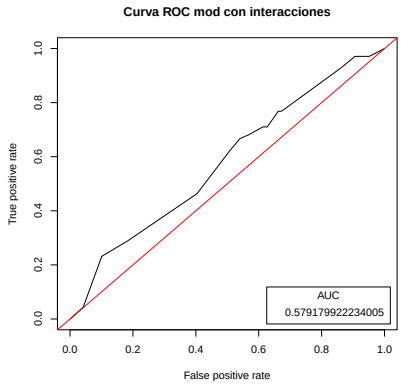


Fig. 3.3: Curva ROC para el modelo de regresión logística con interacciones.

De lo visto en la sección 3.2, en la sección 3.3 y en la tabla 3.2 podemos concluir que ambos modelos son útiles y no hay grandes diferencia de rendimiento, sin embargo, *Naive Bayes* proporciona una facilidad mayor a la hora de interpretar el modelo debido a que se basa en conteos utilizando los datos observados. Por otro lado la regresión logística es un poco más compleja de interpretar y de darle un sentido a los coeficientes calculados, pero no es imposible.

3.5. Conclusiones

Sí se logró predecir la obesidad con un historial de hábitos de ejercicio con un buen porcentaje de aciertos, aunque muy probablemente incluir más variables no tan relacionadas al ejercicio pero sí al perfil de las personas podría mejorar las predicciones.

La aplicación del modelo de Regresión Logística nos mostró resultados muy parecidos a los arrojados por el modelo Naive.

Ha sido de muy enriquecedor poder aplicar dos modelos de clasificación binaria a un mismo conjunto de datos y darse cuenta que los resultados no difieren mucho. Esto habla de la enorme versatilidad que tienen las matemáticas para modelar y resolver un mismo problema. Es interesante notar que al final de cuentas es nuestra creatividad la que nos puede llevar a elegir entre un modelo u otro.

Por otro lado, hay detalles que no pude cubrir. Por ejemplo, me hubiera gustado analizar estos datos por grupos (sexo, edad, grado de estudios, etc.) y ver si la predicción puede mejorar considerándolos. Así como tomar en cuenta más variables que están disponibles pero no consideré a la hora de modelar.

También hay teoría que no pude desarrollar más, por ejemplo, la validación cruzada de los datos para poder entrenar de mejor manera y poder tener un mejor análisis del poder predictivo de ambos modelos. Hizo falta explicar el enfoque bayesiano para construir el score presentado en 3.3. Algunos de estos detalles requerirían de trabajo de cómputo más intenso que sería muy interesante de experimentar en futuros proyectos.

4 Apéndice

4.1. Apéndice Sección 3.2

Ecuación 3.3 se debe a que

$$
\begin{aligned}
S(C|X) &= \ln\left(\frac{P(C|X)}{P(\bar{C}|X)}\right) \\
&= \ln\left(\frac{P(X|C)P(C)}{P(X|\bar{C})P(\bar{C})}\right) \\
&= \ln\left(\frac{P(X|C)}{P(X|\bar{C})}\right) + \ln\left(\frac{P(C)}{P(\bar{C})}\right) \\
&= \ln\left(\frac{\prod_{i=1}^{n} P(X_i|C)}{\prod_{i=1}^{n} P(X_i|\bar{C})}\right) + \ln\left(\frac{P(C)}{P(\bar{C})}\right) \\
&= \sum_{i=1}^{n} \ln\left(\frac{P(X_i|C)}{P(X_i|\bar{C})}\right) + \ln\left(\frac{P(C)}{P(\bar{C})}\right)
\end{aligned}
$$

Bibliografía

[1] Vicente Albíter-Alpízar. “Estudio de patrones de evolución de hábitos en trabajadores de la UNAM, con aplicación en la predicción de obesidad”. Tesis de mtría. Universidad Nacional Autónoma de México, 2021 (vid. págs. 2, 3, 9).

[3] Hedwig Lee, Megan Andrew, Achamyeleh Gebremariam, Julie C. Lumeng y Joyce M. Lee. “Longitudinal Associations Between Poverty and Obesity From Birth Through Adolescence”. En: *American Journal of Public Health* 104.5 (mayo de 2014), e70-e76 (vid. pág. 1).

[5] Juan Rivera Dommarco, ed. *La obesidad en México: estado de la política pública y recomendaciones para su prevención y control*. Primera edición. Mexico: Instituto Nacional de Salud Publica, 2018 (vid. pág. 4).

[6] Juan Rivera Dommarco. *Obesidad en México: recomendaciones para una política de Estado*. 1. ed. México, D.F.: Universidad Nacional Autónoma de México, 2012 (vid. pág. 4).

[7] Romero-Martínez M. Shamah-Levy T. *Encuesta Nacional de Salud y Nutrición 2021 sobre Covid-19. Resultados nacionales.* 1. ed. Cuernavaca, México.: Instituto Nacional de Salud Pública, 2022 (vid. pág. 1).

[8] Christopher R. Stephens, José Antonio Borras Gutiérrez y Hugo Flores. “Bayesian Classification of Personal Histories - An application to the Obesity Epidemic”. en. En: *The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2019)*. Ed. por Aboul Ella Hassanien, Ahmad Taher Azar, Tarek Gaber, Roheet Bhatnagar y Mohamed F. Tolba. Vol. 921. Cham: Springer International Publishing, 2020, págs. 240-249 (vid. págs. 11, 12).

Páginas web

[@2] C3-UNAM. *Proyecto Conductome*. 2022. URL: https://chilam.c3.unam.mx/proyectos/conductome (visitado 22 de sep. de 2022) (vid. pág. 1).

[@4] NHLBI. *Overweight and Obesity - What Are Overweight and Obesity? | NHLBI, NIH*. 2022. URL: https://www.nhlbi.nih.gov/health/overweight-and-obesity (visitado 22 de sep. de 2022) (vid. pág. 1).

[@9] WHO. *Cardiovascular diseases (CVDs)*. en. 2021. URL: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) (visitado 28 de nov. de 2022) (vid. pág. 1).

[@10]WHO. *Physical activity*. 2022. URL: https://www.who.int/news-room/fact-sheets/detail/physical-activity (visitado 7 de nov. de 2022) (vid. pág. 9).