



UNIVERSIDAD NACIONAL DE ENTRE RÍOS
FACULTAD DE INGENIERÍA

**CARRERA: Tecnicatura Universitaria en Procesamiento y
Explotación de Datos**

MATERIA: Modelado Estadístico

Nombre de la Actividad: Trabajo Práctico Final

Fecha de Entrega: 05/06/2023

Profesores:

Mariana Blanco

Juan Aued

Alumnos:

Venturini, Angelo

Ruiz Diaz, Enzo

Introducción.....	3
Exploración.....	5
Hipótesis y preguntas.....	9
Propuestas de modelos predictivos.....	9
Evaluación de los modelos GLM y NB.....	11
Ajuste de los modelos.....	12
Conclusión.....	17

Introducción

Las enfermedades cardiovasculares son una de las principales causas de muerte en todo el mundo. Por esta razón, es fundamental contar con herramientas que permitan detectarlas de manera efectiva. En este contexto, la utilización de técnicas de modelado estadístico se presenta como una herramienta clave para el análisis de datos y la identificación de patrones que permitan predecir la presencia de enfermedades cardiovasculares en pacientes. En este trabajo, se utilizará la base de datos "Cardiovascular Disease dataset", que cuenta con 70000 registros de pacientes y 13 variables, para desarrollar un modelo estadístico que permita predecir la presencia o ausencia de enfermedades cardiovasculares en función de diferentes características de los pacientes. Se analizarán variables objetivas, como la edad, la altura y el peso, así como variables subjetivas, como el consumo de alcohol y la actividad física, entre otras.

La base de datos fue extraída de la siguiente fuente:

Sulianova, K. (2021). Cardiovascular Disease dataset. Recuperado de [Kaggle \(link\)](#)

Hay 3 tipos de características de entrada:

- *Objetivo* : información fáctica
- *Examen* : resultados del examen médico
- *Subjetivo* : información dada por el paciente

Tabla 1*Tabla de variables*

Descripción	Entrada	Identificador	Tipo
Edad	Objetivo	age	int (días)
Altura	Objetivo	height	int(cm)
Peso	Objetivo	weight	float(kg)
Género	Objetivo	gender	1:femenino, 2:masculino
Presión arterial sistólica	Examen	ap_hi	int
Presión arterial diastólica	Examen	ap_lo	int
Colesterol	Examen	cholesterol	1:Normal, 2:Elevado, 3:Muy Elevado
Glucosa	Examen	gluc	1:Normal, 2:Elevado, 3:Muy Elevado
Fumador	Subjetivo	smoke	dicotómica
Consumo de alcohol	Subjetivo	alco	dicotómica
Actividad física	Subjetivo	active	dicotómica
Presencia o ausencia de enfermedad cardiovascular	Variable Objetivo	cardio	dicotómica

Exploración

En esta sección se llevará a cabo una exploración de los datos de la base de datos. El objetivo de esta exploración es comprender mejor la distribución de las variables y detectar posibles valores atípicos o datos faltantes que puedan afectar el análisis posterior. Además, se buscarán posibles correlaciones entre las variables para identificar patrones que puedan ser útiles para la construcción del modelo estadístico.

Se utilizarán herramientas de visualización y estadística descriptiva para llevar a cabo esta exploración.

En primera instancia, se utilizó la función `str()` para constatar los tipos de cada variable con la documentación, también se usó la función `head()` para ver las primeras filas del dataset. Después convertimos la columna `age`, la cual por defecto venía en días, a años. Identificamos que la columna `id` no iba a ser utilizada y procedimos a eliminarla, seguidamente se verificó si había datos null o vacíos. También se detectaron y eliminaron los datos duplicados.

Las variables `cholesterol`, `gender`, `gluc`, `smoke`, `alco`, `cardio` y `active` se pasaron a tipo factor. Además se realizaron las siguientes modificaciones debido a que tenerla en escala numérica no nos resultó intuitivo:

- Cholesterol y gluc: 1 = N, 2 = E y 3 = ME, donde respectivamente refiere a normal, elevado y muy elevado
- Gender: 1 = F (femenino) y 2 = M (masculino)

Realizamos un summary de los datos de las variables cuantitativas.

Gráfico 1

Estimadores brindados por la función summary

age	height	weight	ap_hi	ap_lo
Min. :29.00	Min. : 55.0	Min. : 10.00	Min. : -150.0	Min. : -70.00
1st Qu.:48.00	1st Qu.:159.0	1st Qu.: 65.00	1st Qu.: 120.0	1st Qu.: 80.00
Median :53.00	Median :165.0	Median : 72.00	Median : 120.0	Median : 80.00
Mean :52.86	Mean :164.3	Mean : 74.52	Mean : 129.2	Mean : 97.45
3rd Qu.:58.00	3rd Qu.:170.0	3rd Qu.: 83.00	3rd Qu.: 140.0	3rd Qu.: 90.00
Max. :64.00	Max. :250.0	Max. :200.00	Max. :16020.0	Max. :11000.00

Observamos en el gráfico 1, que algunas variables presentan irregularidades en sus máximos y mínimos, como por ejemplo, `ap_hi` que hace referencia a la presión sistólica del paciente, presenta valores negativos y excesivamente grandes lo cual es erróneo. De forma similar ocurre con la presión diastólica, la altura y el peso. Dada esta situación el grupo decidió hacer un recorte del 2,5% de los datos para los extremos de las variables, para de esta forma eliminar los outliers que fueron presentados.

Adicionalmente, se realizó la verificación de que la presión diastólica sea menor a la presión

sistólica para cada paciente, ya que es una condición imposible de no cumplirse en un paciente.

Luego de realizadas las adecuaciones anteriormente mencionadas, volvimos a realizar un summary.

Gráfico 2

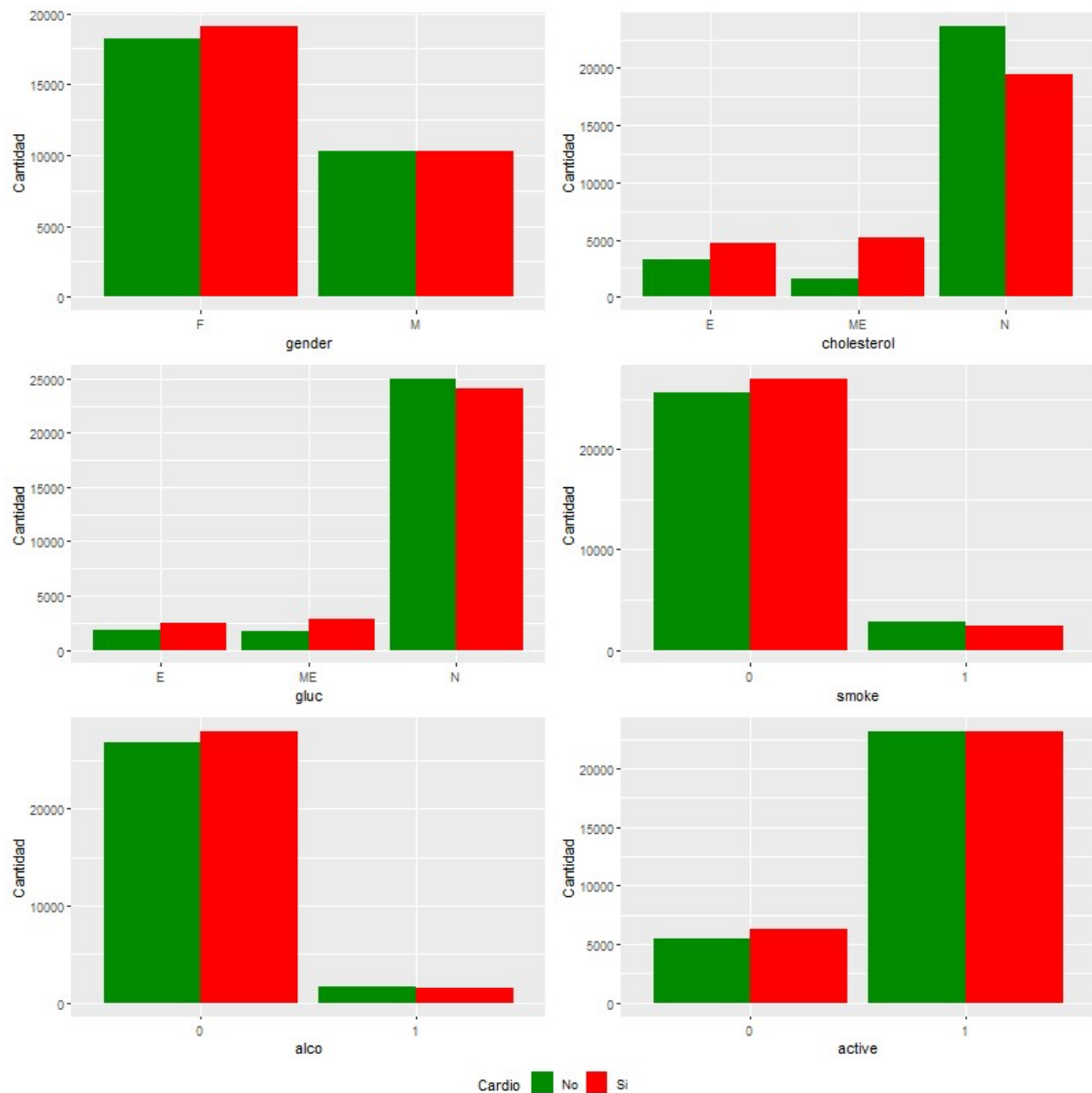
Estimadores de las variables posteriores a la limpieza

age	height	weight	ap_hi	ap_lo
Min. :29.00	Min. :150.0	Min. : 52.00	Min. :100.0	Min. : 60.00
1st Qu.:48.00	1st Qu.:160.0	1st Qu.: 65.00	1st Qu.:120.0	1st Qu.: 80.00
Median :54.00	Median :165.0	Median : 72.00	Median :120.0	Median : 80.00
Mean :52.89	Mean :164.6	Mean : 73.95	Mean :126.6	Mean : 81.37
3rd Qu.:58.00	3rd Qu.:170.0	3rd Qu.: 82.00	3rd Qu.:140.0	3rd Qu.: 90.00
Max. :64.00	Max. :180.0	Max. :108.00	Max. :170.0	Max. :110.00

Luego del análisis estadístico de nuestras variables cuantitativas, procederemos a ver la distribución de nuestras variables cualitativas con respecto a cardio, que sería nuestra variable a objetivo.

Gráfico 3

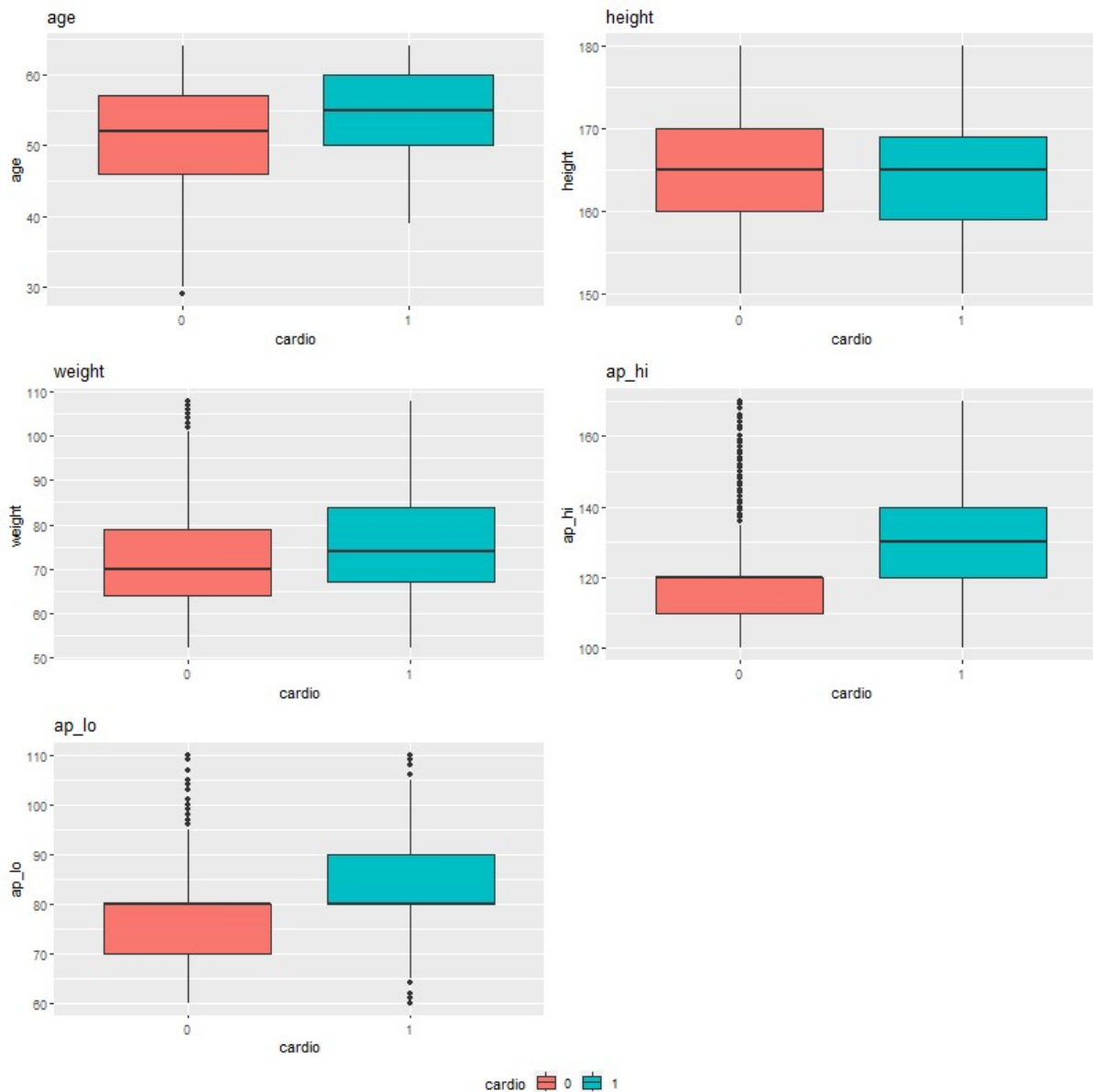
Gráfico de barras para cada variable cualitativa con respecto a cardio



Ahora bien, podríamos hacer suposiciones con respecto al gráfico 3, por ejemplo vemos que para la variable colesterol, la cantidad de pacientes enfermos es mayor para los que tienen el colesterol muy elevado. Para la variable gluc, no es tan marcada esa diferencia como en colesterol pero se podría decir que hay una relación. En cambio con los demás, no parece haber una relación directa.

Gráfico 4

Gráfico de cajas y bigotes para las variables cuantitativas con respecto a cardio



Acto seguido, veremos cómo se comportan nuestras variables cuantitativas con respecto a cardio mediante gráficos de boxplot. En el gráfico 4, podemos notar que en el caso de los pacientes que presentan una enfermedad cardiovascular son de mayor edad, tienen mayor peso y mayor presión tanto sistólica como diastólica, por lo que podemos suponer que entre la variable cardio y las anteriormente mencionadas hay una posible relación.

Planteo de preguntas

En esta sección se plantearán preguntas que guiarán el análisis del modelo estadístico a desarrollar. Estas preguntas surgirán a partir de la exploración previa de los datos y se utilizarán para definir las variables a incluir en el modelo y para evaluar su capacidad predictiva.

Todo lo anteriormente mencionado nos lleva a plantear lo siguiente:

¿Qué variables de la base de datos son buenos indicadores independientes de la presencia de enfermedad cardiovascular?

¿Hay variables que puedan ser excluidas del modelo predictivo?

¿Qué modelo de los propuestos predice mejor la presencia de enfermedades cardiovasculares en los pacientes?

Propuestas de modelos predictivos

Para intentar responder las preguntas planteadas, el grupo propuso dos modelos estadísticos. En primer lugar se propone el modelo Naive-Bayes debido a la presencia de múltiples variables categóricas, y la ventaja que este modelo presenta ante la gran cantidad de datos es su rapidez y simplicidad. Por otro lado, se propone utilizar el modelo de regresión logística debido a que es adecuado para grandes volúmenes de datos y presenta la ventaja de ser flexible y fácil de interpretar..

Para mejorar el rendimiento de los modelos, se utilizarán variables dummies para representar las variables categóricas en los modelos y definir el efecto de las mismas sobre la predicción.

Además, se utilizará la estrategia de K-Fold cross validation para obtener una estimación más confiable del rendimiento de los modelos y utilizar de forma más eficiente los datos en el entrenamiento, específicamente 10–Fold cross validation. En resumen, se espera que estos modelos permitan identificar las variables más importantes para predecir la presencia de enfermedad cardiovascular en un paciente.

Evaluación de los modelos GLM y NB

En esta sección, se llevará a cabo la evaluación de los modelos estadísticos propuestos previamente, el modelo Bayes-Ingenuo y el modelo de regresión logística. Para ello, se utilizarán herramientas como la curva ROC y la precisión (accuracy) para determinar la efectividad de cada modelo en la predicción de la presencia de enfermedades cardiovasculares en pacientes.

Para el modelo de Bayes Ingenuo obtuvimos un área bajo la curva ROC de 0,6785 y un accuracy de 0,6771 y para el modelo de regresión logística obtuvimos un área bajo la curva ROC de 0,72 y un accuracy de 0,7201 , como podemos observar en el gráfico 6. Además, en el gráfico 5 vemos que ambas curvas ROC son buenas, destacando más el modelo de regresión logística.

Gráfico 5

Curvas ROC de los modelos Bayes Ingenuo y Regresión Logística

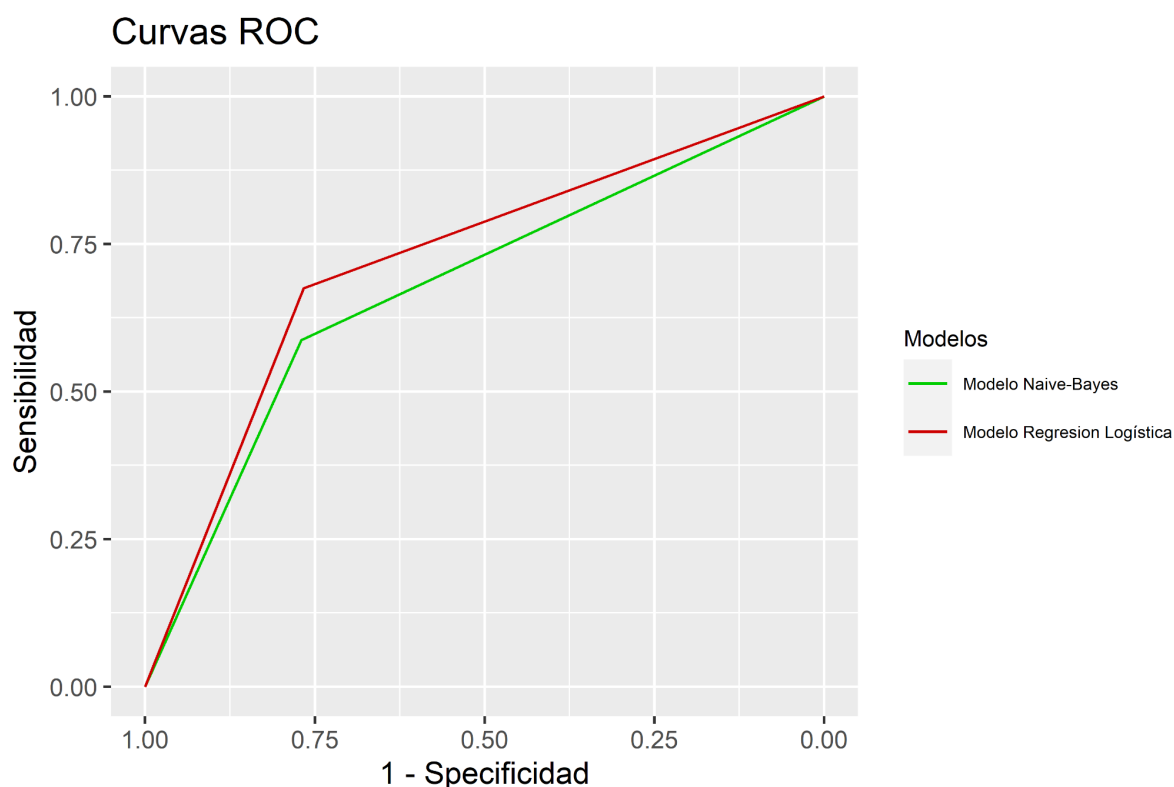


Gráfico 6

Valores del área bajo la curva ROC y accuracy (precisión) para cada modelo

```
Valores de referencia:  
Modelo Naive-Bayes:  
Valor AUC: 0.6785  
Accuracy: 0.6771599  
Modelo Regresión Logística:  
Valor AUC: 0.7208  
Accuracy: 0.7201167
```

Gráfico 7

Matriz de confusión para el modelo de Naive-Bayes

		Reference	
Prediction		0	1
	0	21969	12141
	1	6565	17267

Gráfico 8

Matriz de confusión para el modelo de Regresión Logística

		Reference	
Prediction		0	1
	0	21877	9560
	1	6657	19848

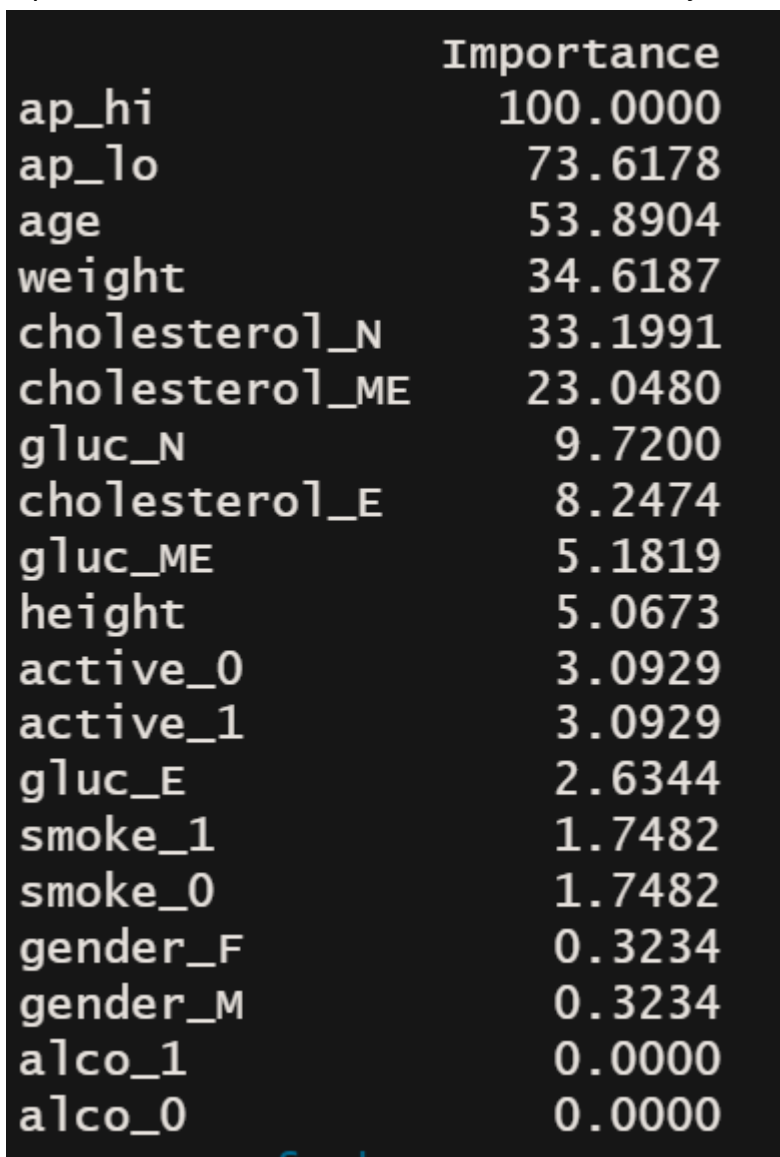
Se observa en el gráfico 7 que el modelo de Naive-Bayes es moderadamente sensible y específico. A su vez en el gráfico 8 se puede observar que el modelo de regresión logística es más específico pero la sensibilidad es casi idéntica.

Ajuste de los modelos

En esta sección, nos enfocaremos en el ajuste de los modelos de regresión logística y Naive-Bayes. Para seleccionar las variables significativas utilizaremos los p-valor en el caso de la regresión logística y para el modelo de Naive-Bayes usaremos la función varImp del paquete caret y de esta manera poder ajustar los modelos de manera óptima, utilizando solamente variables que sean significativas. El análisis de los modelos resultantes de estas modificaciones nos permitirá identificar las variables más relevantes en la predicción de la enfermedad y evaluar la eficacia de los nuevos modelos ajustados para seleccionar el más óptimo.

Gráfico 9

Importancia de las variables del modelo de Naive-Bayes



	Importance
ap_hi	100.0000
ap_lo	73.6178
age	53.8904
weight	34.6187
cholesterol_N	33.1991
cholesterol_ME	23.0480
gluc_N	9.7200
cholesterol_E	8.2474
gluc_ME	5.1819
height	5.0673
active_0	3.0929
active_1	3.0929
gluc_E	2.6344
smoke_1	1.7482
smoke_0	1.7482
gender_F	0.3234
gender_M	0.3234
alco_1	0.0000
alco_0	0.0000

Para la evaluación de las variables de importancia para el modelo de Naive-Bayes, se ordenan las mismas con respecto a la importancia en la curva ROC en una escala del 0 al 100. Como podemos observar en el gráfico 9, y tomando un valor arbitrario mayor a 5 para decidir si una variable es importante o no, que las variables significativas para el modelo son: ap_hi, ap_lo, age, weight, cholesterol_N, cholesterol_ME, gluc_N, cholesterol_E, gluc_ME, height. A continuación, procederemos a generar nuevamente el modelo, pero solo con las variables significativas.

Gráfico 10

Curvas ROC de los modelos Naive-Bayes y Naive-Bayes ajustado

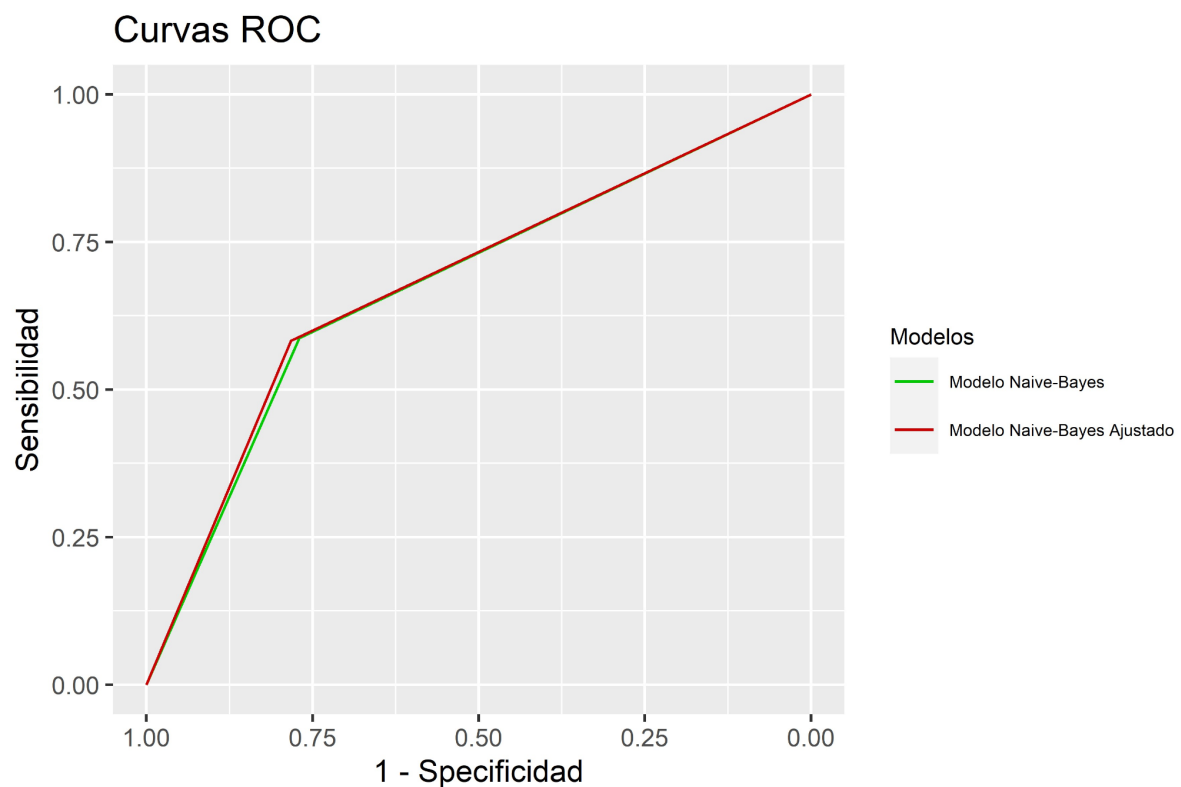
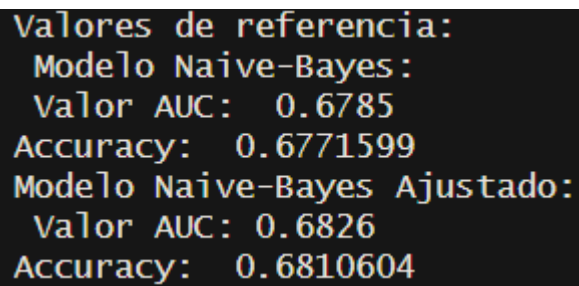


Gráfico 11

Valores del área bajo la curva ROC y accuracy (precisión) para cada modelo



```
Valores de referencia:  
Modelo Naive-Bayes:  
Valor AUC: 0.6785  
Accuracy: 0.6771599  
Modelo Naive-Bayes Ajustado:  
Valor AUC: 0.6826  
Accuracy: 0.6810604
```

Observando el gráfico 10 y 11, podemos ver una leve mejoría en los resultados del modelo, por lo que podemos decir que el modelo mejoró con las variables elegidas, pero que no fue muy significativo el cambio entre ellos. De hecho, ahora el modelo es levemente más sensible.

Para la evaluación de las variables de mayor significancia en el modelo de regresión logística, observando los p-valores y las estrellas de significancia, y tomando como significativo el valor más alejado a 0.05 en el p-value y la cantidad de estrellas mayor a 1, vemos que las variables significativas son: age, weight, ap_hi, ap_lo, cholesterol_E, cholesterol_ME, gluc_ME, smoke_0, alco_0 y active_0. Esto se puede ver en mayor detalle en el gráfico 12.

Gráfico 12

Coeficientes del modelo de Regresión Logística (summary)

```
coefficients: (6 not defined because of singularities)
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.201e+01  3.108e-01 -38.626  < 2e-16 ***
age          4.942e-02  1.456e-03  33.940  < 2e-16 ***
height      -3.223e-03  1.639e-03  -1.967   0.0492 *
weight       1.013e-02  8.598e-04  11.777  < 2e-16 ***
ap_hi        6.071e-02  1.039e-03  58.441  < 2e-16 ***
ap_lo        1.200e-02  1.604e-03   7.476  7.65e-14 ***
gender_F     4.461e-02  2.388e-02   1.868   0.0617 .
gender_M          NA          NA      NA      NA
cholesterol_E 3.352e-01  2.931e-02  11.437  < 2e-16 ***
cholesterol_ME 1.065e+00  3.821e-02  27.862  < 2e-16 ***
cholesterol_N          NA          NA      NA      NA
gluc_E       -1.555e-02  3.866e-02  -0.402   0.6875
gluc_ME      -3.404e-01  4.194e-02  -8.116  4.80e-16 ***
gluc_N          NA          NA      NA      NA
smoke_0       1.817e-01  3.715e-02   4.892  9.99e-07 ***
smoke_1          NA          NA      NA      NA
alco_0        2.309e-01  4.496e-02   5.135  2.82e-07 ***
alco_1          NA          NA      NA      NA
active_0       2.048e-01  2.345e-02   8.736  < 2e-16 ***
active_1          NA          NA      NA      NA
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Se generó nuevamente el modelo con las variables anteriormente mencionadas y los resultados fueron los siguientes:

Gráfico 13

Curvas ROC de los modelos de Regresión Logística y Regresión Logística Ajustada

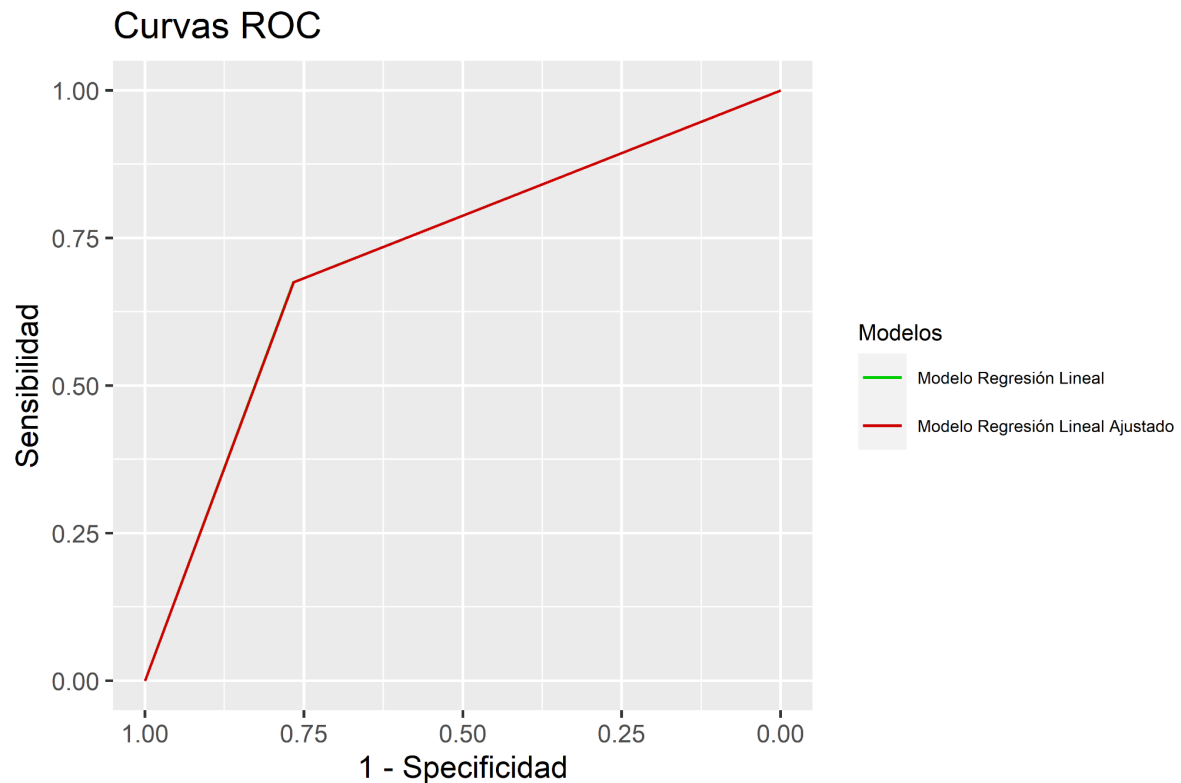


Gráfico 14

Valores del área bajo la curva ROC y accuracy (precisión) para cada modelo

```
Valores de referencia:  
Modelo Regresión Logística:  
Valor AUC: 0.7208  
Accuracy: 0.7201167  
Modelo Regresión Logística Ajustado:  
Valor AUC: 0.7207  
Accuracy: 0.7199613
```

Al contrario del caso anterior, como podemos observar en el gráfico 13 y 14, el modelo no mejoró, notando un muy leve descenso del accuracy. Es más, se podría decir que es casi idéntico, por lo tanto podemos decir que el modelo inicial ya estaba ajustado.

Conclusión

En conclusión, después de evaluar los diferentes modelos propuestos, podemos afirmar que el modelo de regresión logística es el que mejor predice la presencia de una enfermedad cardiovascular en los pacientes. Este modelo ha demostrado ser el más efectivo en términos de precisión y fiabilidad, lo que lo convierte en la mejor opción para predecir los resultados de la variable dependiente.

En términos de precisión y fiabilidad, se puede afirmar que las variables que fueron declaradas como no significativas en el modelo ajustado pueden ser excluidas sin generar una diferencia significativa en la predicción de la variable dependiente.

También es posible afirmar que las variables de edad, peso, presión arterial sistólica, presión arterial diastólica, niveles elevados y muy elevados de colesterol, niveles muy elevados de glucosa, el hábito de fumar, el consumo de alcohol y la actividad física son significativas para el modelo.