Tomas Uzquiano

Enfermedades cardíacas

Trabajo final
Comisión 32775

ÍNDICE

Abstract

Las enfermedades y los problemas del corazón afectan a una gran parte de la sociedad. Nos afectan a nosotros, ya sea de forma directa o a nuestros familiares y amigos. Un accidente cerebrovascular puede causar ocasionalmente discapacidades temporales o permanentes, según el tiempo que el cerebro permanezca sin flujo sanguíneo y la parte afectada.
Muchas estrategias de prevención de accidentes cerebrovasculares son las mismas que las estrategias de prevención de enfermedades cardíacas. Controlar la presión arterial alta, reducir la cantidad de colesterol y grasas saturadas en tu alimentación, hacer ejercicio de forma regular, evitar el alcohol y el tabaquismo entre otras recomendaciones para mantener un estilo de vida saludable.
Es esencial la detección temprana de los distintos signos de advertencia para evitar o minimizar un accidente cerebrovascular. Se busca detectar en un estadío temprano cualquier tipo de enfermedad por accidente cerebrovascular a través del uso de un modelo creado a partir de una base de datos de miles de personas con múltiples características para cada una. Este dataset pertenece al BRFSS 2015 (Behavioral Risk Factor Surveillance System).

Objetivo

a. ¿Cómo se puede prevenir la ocurrencia de un ataque al corazón?
b. ¿Cómo se puede prevenir la ocurrencia de un accidente cerebrovascular?
c. ¿Cómo la presencia de hábitos saludables y de una alimentación adecuada afectan a la ocurrencia de accidentes cerebrovasculares?

Hipótesis

a. ¿Qué factores están mayormente relacionados con la ocurrencia de un ataque al corazón?
b. ¿Puede la existencia de un ataque al corazón llevar a una persona a sufrir de un accidente cerebrovascular?

Adquisición de datos

El Behavioral Risk Factor Surveillance System (BRFSS) es un proyecto de colaboración entre todos los estados de los Estados Unidos (EE. UU.) y los territorios participantes de los Estados Unidos y los Centers for Disease Control and Prevention (CDC). El dataset posee 253,680 filas y 22 columnas pertenecientes a las características obtenidas de respuestas de la encuesta de BRFSS 2015 - clasificación binaria. 229,787 encuestados no tienen / no han tenido enfermedades cardíacas, mientras que 23,893 han tenido enfermedad cardíaca. Si quiere investigar más sobre este dataset ingrese a:
https://www.cdc.gov/brfss/annual_data/2015/pdf/codebook15_llcp.pdf

Diccionario de variables

* HeartDiseaseorAttack: Ataque al corazón o enfermedad al corazón. 1 significa que la persona presenta problemas o sufrió de un ataque al corazón. Variable binaria.

* HighBP: Alta presión sanguínea. 1 significa que la persona ha sido indicada por un profesional de la salud que presenta una presión sanguínea elevada. Variable binaria.

* HighChol: Alto colesterol. 1 significa que la persona ha sido indicada por un profesional de la salud que presenta alto colesterol. Variable binaria.

* CholCheck: Control de colesterol. 1 significa que la persona se ha hecho un control de colesterol. Variable binaria.

* BMI: Índice de masa corporal (IMC). Variable entera entre 0 y 100.

* Smoker: Fumador. 1 Significa que la persona ha fumado al menos 100 cigarrillos en su vida. Variable binaria.

* Stroke: Accidente cerebrovascular. 1 significa que a la persona le han dicho que sufrió de un accidente cerebrovascular. Variable binaria.

* Diabetes: Diabetes. 1 significa que a la persona le han dicho que sufrió de diabetes. Variable binaria.

* Physical Activity: Actividad física. 1 si la persona reportó haber hecho ejercicio en los ultimos 30 dias. Variable binaria.

* Fruits: Frutas. 1 si la persona consume 1 o más frutas por día. Variable binaria.

* Vegetables: Verduras. 1 si la persona consume 1 o más verduras por día. Variable binaria.

* HvyAlcoholConsump: Consumo excesivo de alcohol. 1 si la persona adulta: es hombre y consumió más de 14 vasos de alcohol en la semana o si es mujer y consumió más de 7 vasos de alcohol en la semana. Variable binaria.

* AnyHealthcare: Seguro social. 1 si la persona tiene algún tipo de plan de salud, ya sea seguro social, plan prepago o plan de seguro gubernamental. Variable binaria.

* NoDocbcCost: No le fue posible atenderse con un doctor por falta de dinero. 1 si la persona se quiso atender con un doctor y no pudo por falta de dinero. Variable binaria.

* GenHlth: 5 diferentes categorías que indican el estado de salud general.
  1. Excelente
  2. Muy bueno
  3. Bueno
  4. Aceptable
  5. Malo

* MentHlth: Salud mental. Pensando sobre salud mental, la cual incluye estrés, depresión y problemas con tus emociones. ¿Por cuántos días durante los últimos 30 días tu salud mental no ha sido buena? Variable entera entre 0 y 30.

* PhysHlth: Salud física. Pensando en tu salud física, la cual incluye enfermedades y lesiones físicas, ¿por cuántos días durante los últimos 30 días tu salud física no ha sido buena? Variable entera entre 0 y 30.

* DiffWalk: Dificultad para caminar. 1 si la persona tiene problemas para caminar o subir escaleras. Variable binaria

* Sex: Indica el sexo de la persona. 0 si la persona es mujer y si la persona es hombre. Variable binaria.

* Age: 14 diferentes categorías que indican la edad.

  1. 18 <= Edad <= 24
  2. 25 <= Edad <= 29
  3. 30 <= Edad <= 34
  4. 35 <= Edad <= 39
  5. 40 <= Edad <= 44
  6. 45 <= Edad <= 49
  7. 50 <= Edad <= 54
  8. 55 <= Edad <= 59
  9. 60 <= Edad <= 64
  10. 65 <= Edad <= 69
  11. 70 <= Edad <= 74
  12. 75 <= Edad <= 79
  13. 80 <= Edad <= 99

* Education: 6 diferentes categorías que indican la educación. Indica el mayor de nivel de educación completado.

1. Solo atendió al jardín de infantes/preescolar
2. Grados entre 1-8. Educación elemental
3. Grados entre 9-11. Educación superior/secundaria en curso
4. Grados 12. Educación superior/secundaria completa
5. Universidad 1-3 años. Tecnicatura completa o universidad en curso
6. Universidad 4 años o más. Universidad completa

* Income: 8 diferentes categorías que indican los ingresos anuales en dólares por persona.

1. Ingresos <= $10000
2. $10000 < Ingresos <= $15000
3. $15000 < Ingresos <= $20000
4. $20000 < Ingresos <= $25000
5. $25000 < Ingresos <= $35000
6. $35000 < Ingresos <= $50000
7. $50000 < Ingresos <= $75000
8. $75000 < Ingresos

Limpieza de datos

Se buscó la aparición de algún dato nulo en el dataset. No se encontró ninguno.

```
HeartDiseaseorAttack    False
HighBP                  False
HighChol                False
CholCheck               False
BMI                     False
Smoker                  False
Stroke                  False
Diabetes                False
PhysActivity            False
Fruits                  False
Veggies                 False
HvyAlcoholConsump       False
AnyHealthcare           False
NoDocbcCost             False
GenHlth                 False
MentHlth                False
PhysHlth                False
DiffWalk                False
Sex                     False
Age                     False
Education               False
Income                  False
dtype: bool
```

Se verificó que no haya datos faltantes. Todo resultó correcto.

```
HeartDiseaseorAttack    0
HighBP                  0
HighChol                0
CholCheck               0
BMI                     0
Smoker                  0
Stroke                  0
Diabetes                0
PhysActivity            0
Fruits                  0
Veggies                 0
HvyAlcoholConsump       0
AnyHealthcare           0
NoDocbcCost             0
GenHlth                 0
MentHlth                0
PhysHlth                0
DiffWalk                0
Sex                     0
Age                     0
Education               0
Income                  0
dtype: int64
```

Se renombró a las variables con nombres más claros, significativos y en español.

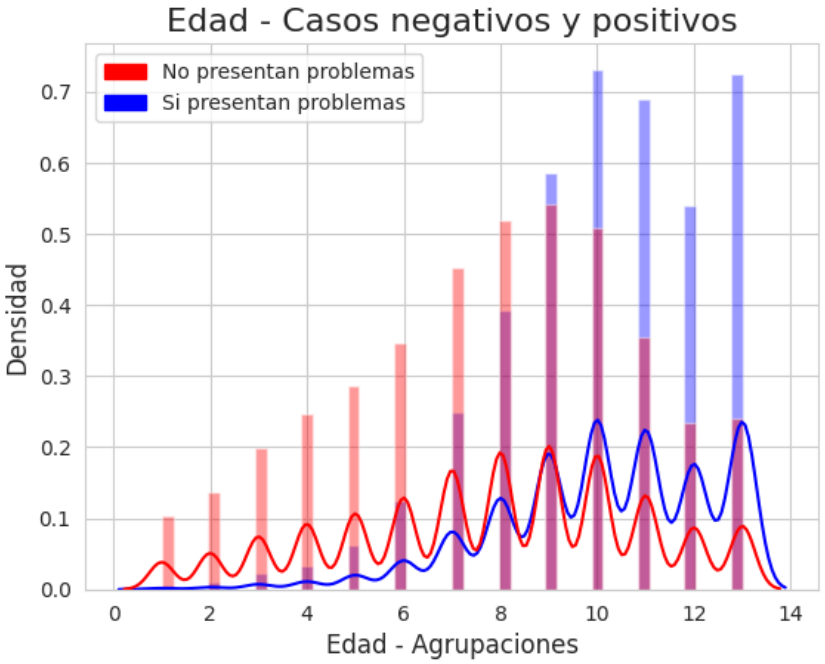
Nombre original	Nombre modificado
HeartDiseaseorAttack	target
HighBP	AltaPresionSanguinea
HighChol	AltoColesterol
CholCheck	Colesterol
Smoker	Fumador
Stroke	AccidenteCerebrovascular
PhysActivity	ActivFisica
Fruits	ConsumoFrutas
Veggies	ConsumoVerduras
HvyAlcoholConsump	AltoConsumoAlcohol
AnyHealthcare	ObraSocial
NoDocbcCost	SinDoctorCostoso
GenHlth	SaludGnrl
DiffWalk	DificCaminar
Sex	Genero
Education	Educacion
Income	IngresoMonetario
Age	Edad
BMI	IMC
MentHlth	SaludMental
PhysHlth	SaludFisica

Se decidió utilizar Min-Max Scaler para normalizar los datos dentro del rango 0 a 1. De esta forma se pudo mejorar la precisión y el rendimiento de los modelos de aprendizaje automático teniendo en cuenta que utilizamos datos en diferentes escalas y unidades de medida. Esto fue aplicado a las variables:

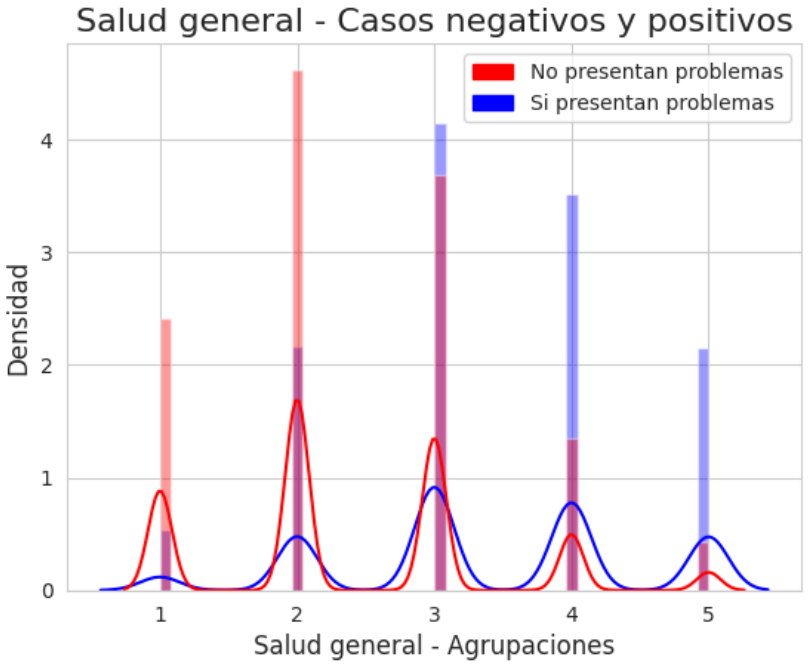
a. IMC
b. Edad
c. SaludFisica
d. SaludMental

Análisis de datos

Se comenzó con un análisis introductorio de algunas características para casos positivos y negativos de enfermedades cardíacas.

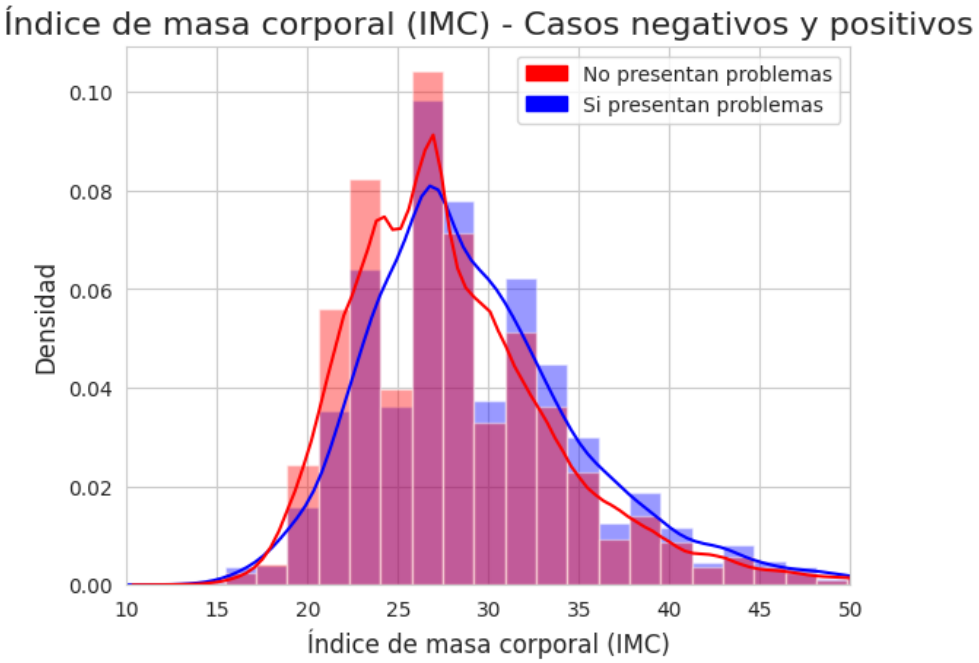


Se pudo observar que aquellas personas con presencia de enfermedades cardíacas eran de una edad mayor que aquellas que no presentaban ningún síntoma.

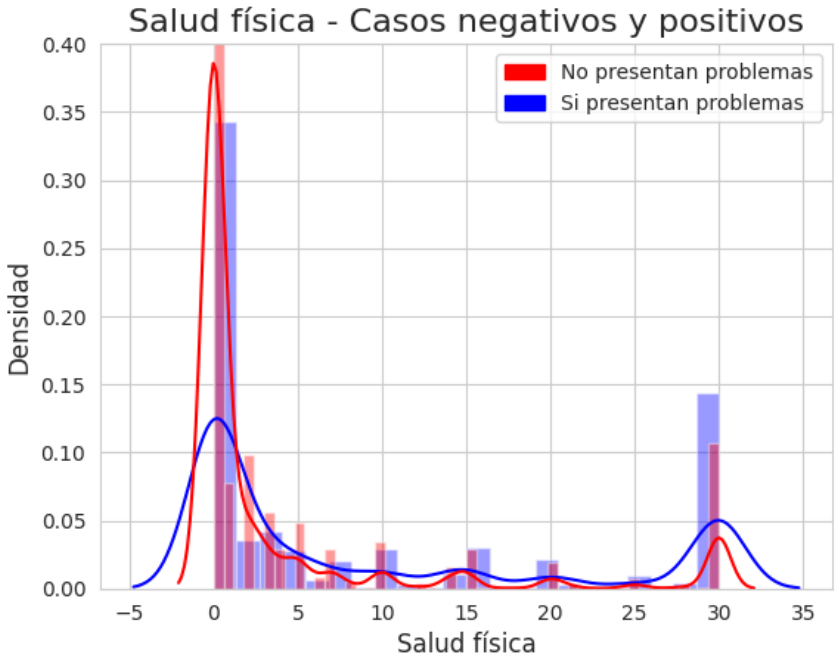


Se pudo observar que aquellas personas con presencia de enfermedades cardíacas tenían un estado de salud general mejor que aquellas que no presentaban ningún síntoma. Se

observó que en 1,2,3 la densidad de personas que no presentan síntomas es mayor. Por otra parte, en 4,5 las personas que presentan síntomas es mayor.



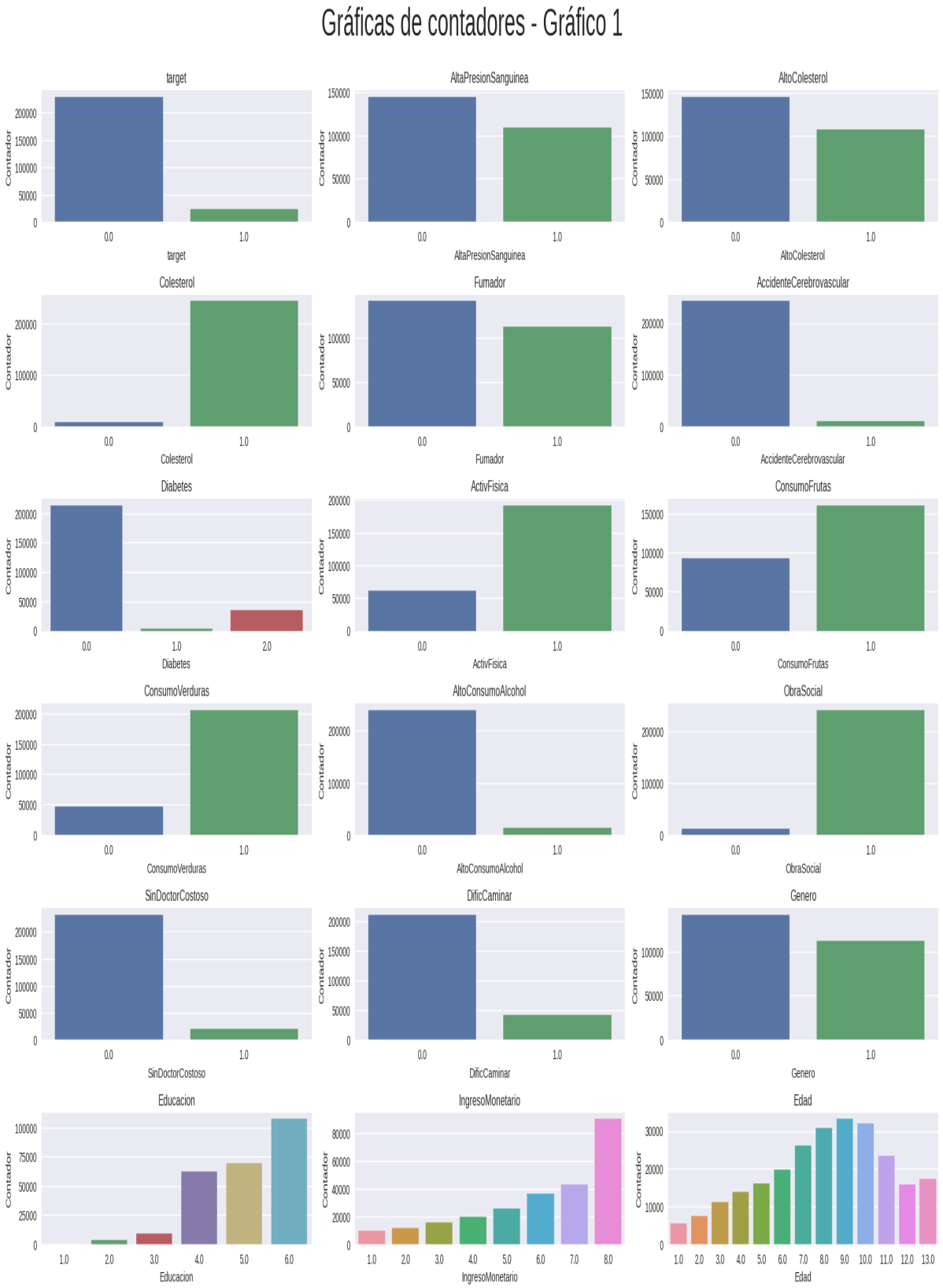
Se pudo observar que aquellas personas con presencia de enfermedades cardíacas tenían un IMC menor que aquellas que no presentaban ningún síntoma.



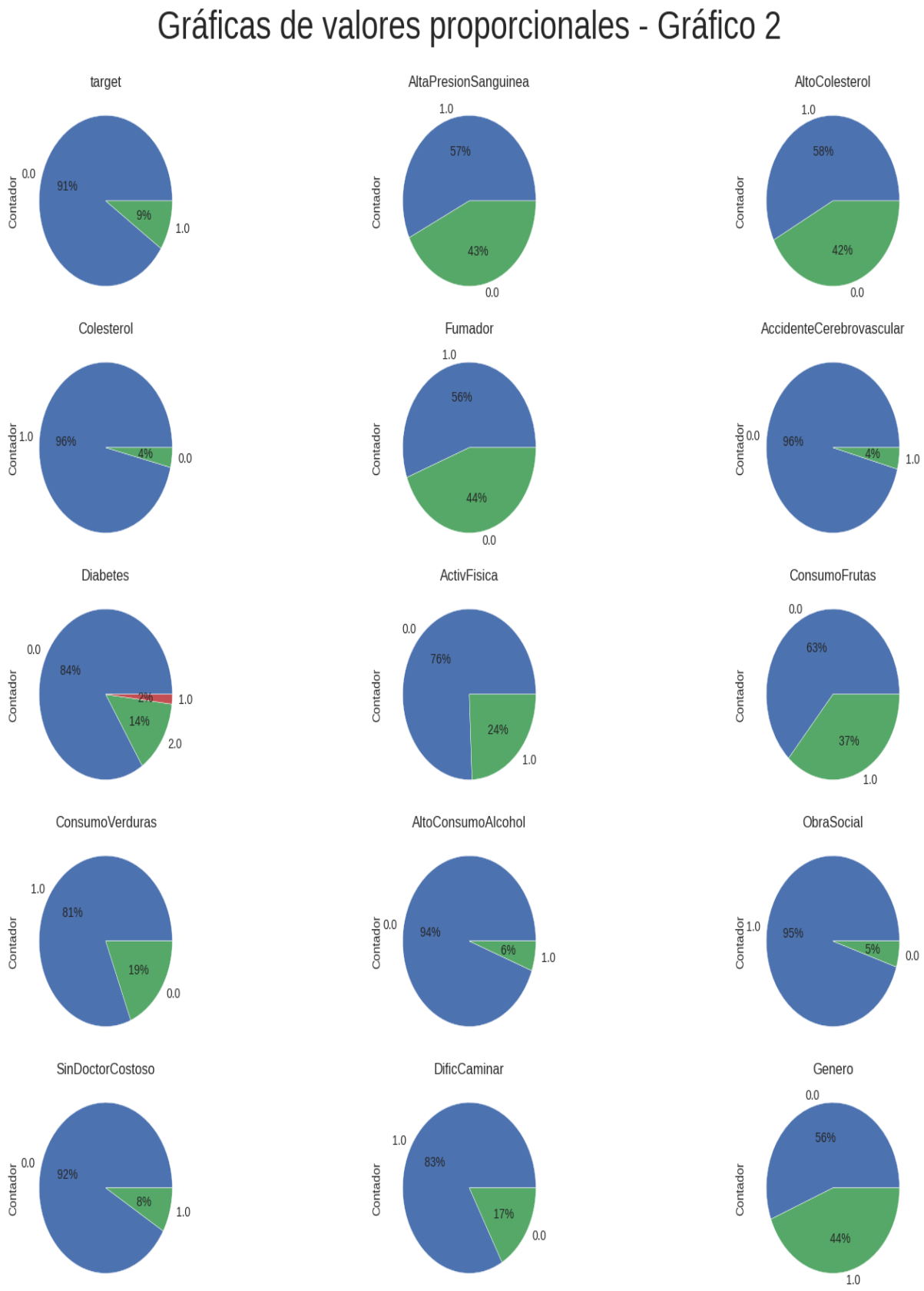
Para recordar, la variable analizada representa la cantidad de días de los últimos 30 días en los cuales su salud física no ha sido buena. Por lo que se pudo observar que aquellas personas con presencia de enfermedades cardíacas tenían peor salud física que aquellas que no presentaban ningún síntoma.

Primer gráfico inmersivo

Los siguientes tres gráficos inmersivos fueron representados con el objetivo de obtener algunos insights generales y poder profundizar el análisis como se verá a partir del gráfico 4.



Segundo gráfico inmersivo

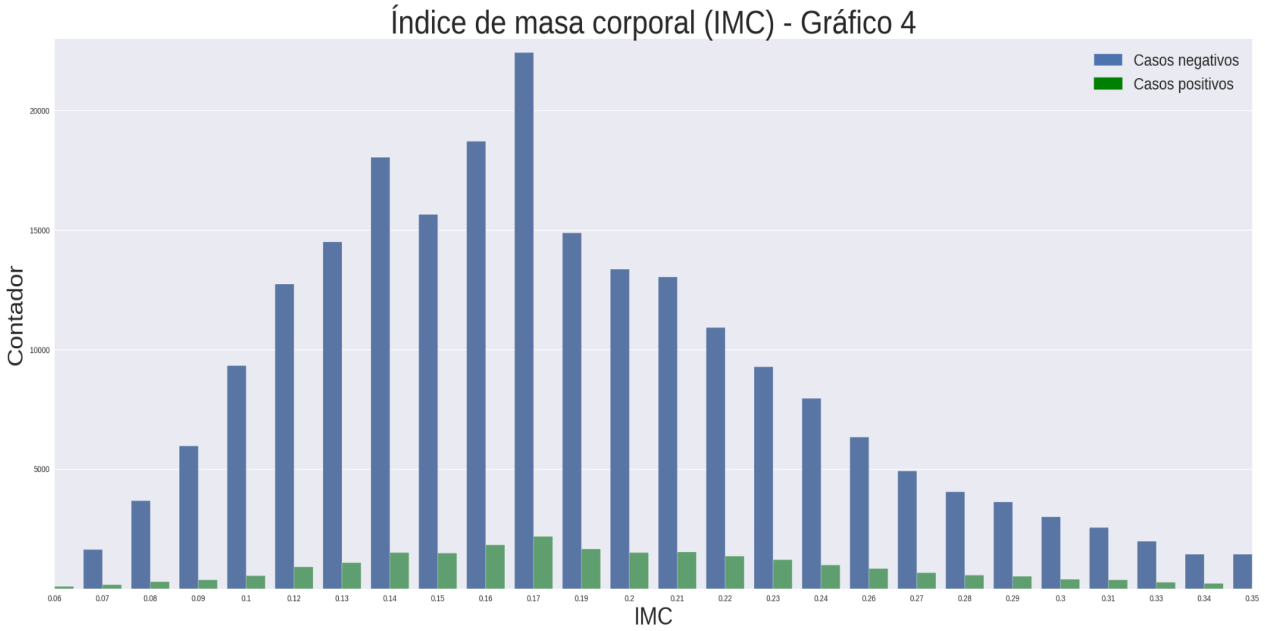


Se observó en el "Gráfico 2" que la variable target presentó mayormente ausencia de problemas en el corazón (91%).
No se vió un problema importante de accidentes cerebrovasculares (4% positivo), gran consumo de alcohol (6% positivo) o problemas al caminar (17% positivo). Además, la mitad de la prueba aproximadamente presentó:

1. Alto colesterol (58%).
2. Alta presión en sangre (43%).
3. Fumadores activos (56%).
4. Sexo masculino (56%).

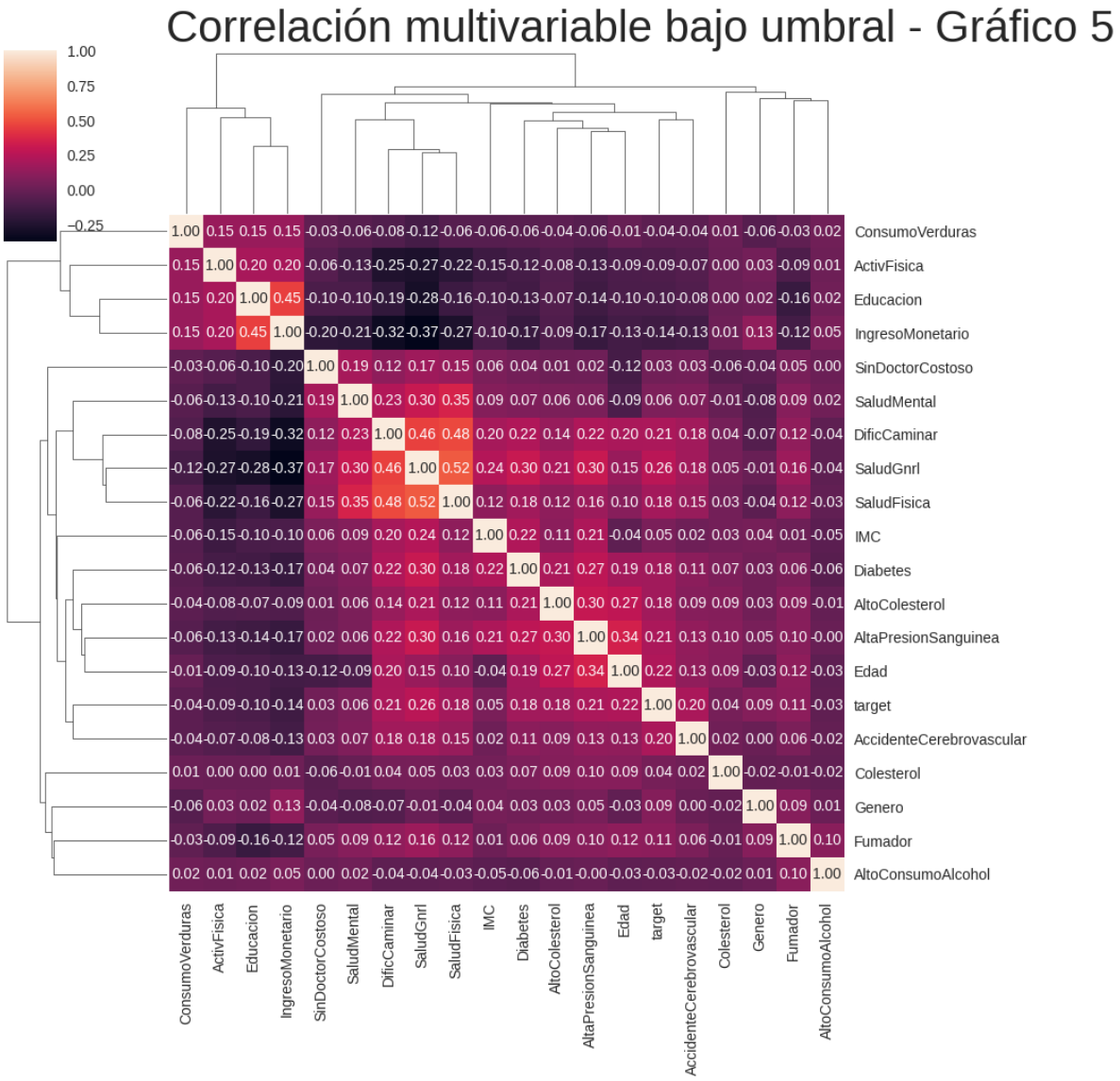
Tercer gráfico inmersivo

Gráfico 4



Se observó en el "Gráfico 4" una leve correlación entre el índice de masa corporal y la presencia de problemas en el corazón. En presencia y ausencia de problemas en el corazón se observó una distribución similar.

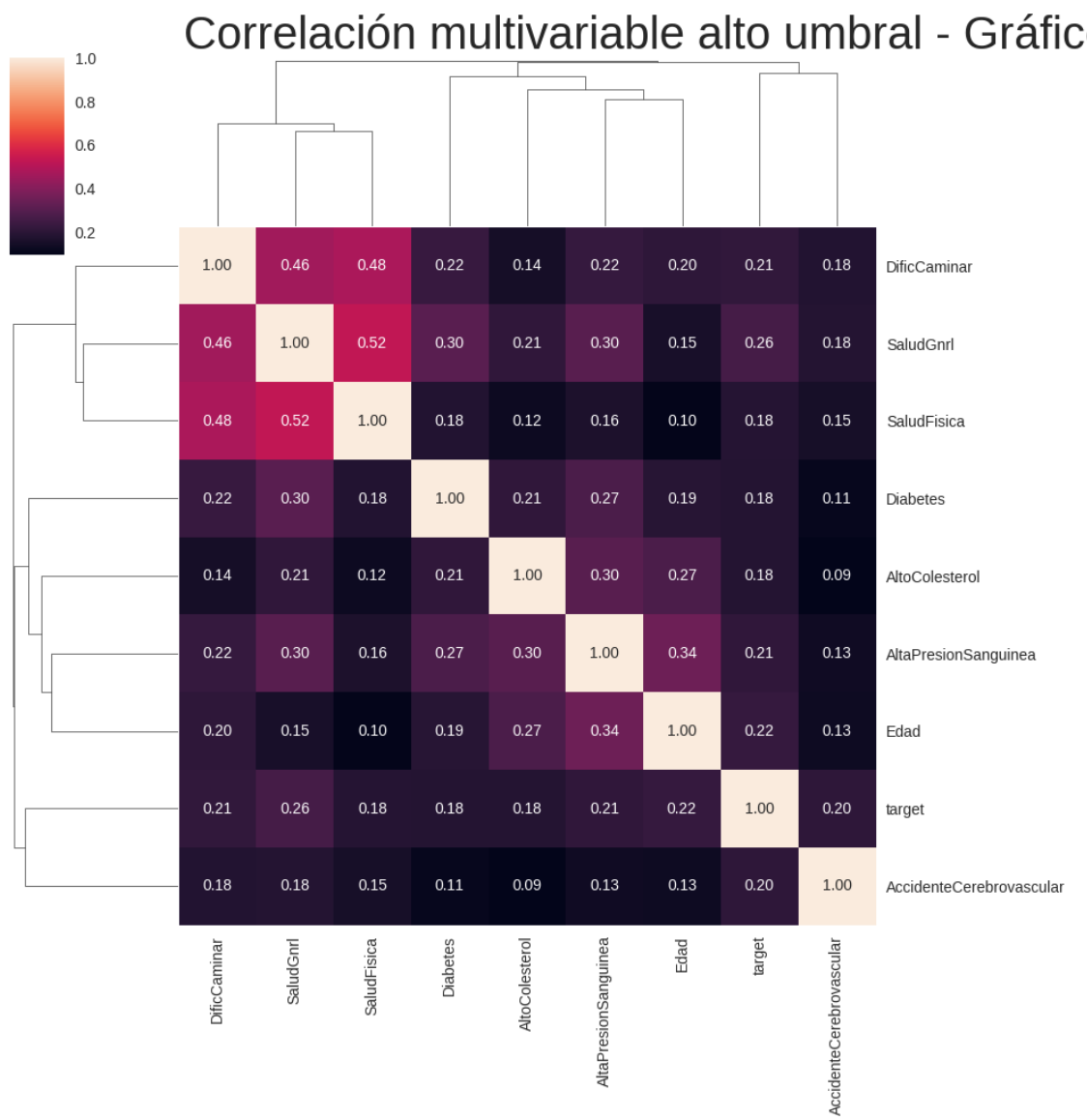
Gráfico 5



Se observó en el gráfico 5" una matriz de relación simétrica. Es por esta razón que solo se necesitó la parte superior a la diagonal principal para sacar algunas conclusiones. Utilizando un umbral bajo (0.025) se observó que:

1. DificCaminar, (dificultades al caminar) tiene una correlación fuerte con SaludGnrl (salud general) 0.46 y PhysHlth (salud física) 0.48, así como SaludGnrl con SaludFisica (salud fisica) 0.52.
2. IngresoMonetario parece ser poco útil debido a que no tiene correlación con las demás características. Pudo ser una opción eliminarlo del dataframe.
3. Nuestro target esta poco correlacionado con el IMC (índice de masa corporal) 0.05, SaludMental (0.06) y SinDoctorCostoso (0.03).
4. Nuestro target encontró una correlación negativa con IngresoMonetario (-0.14), Educacion (-0.1), ConsumoVerduras (-0.04) y ActivFisica (-0.09).

Gráfico 6

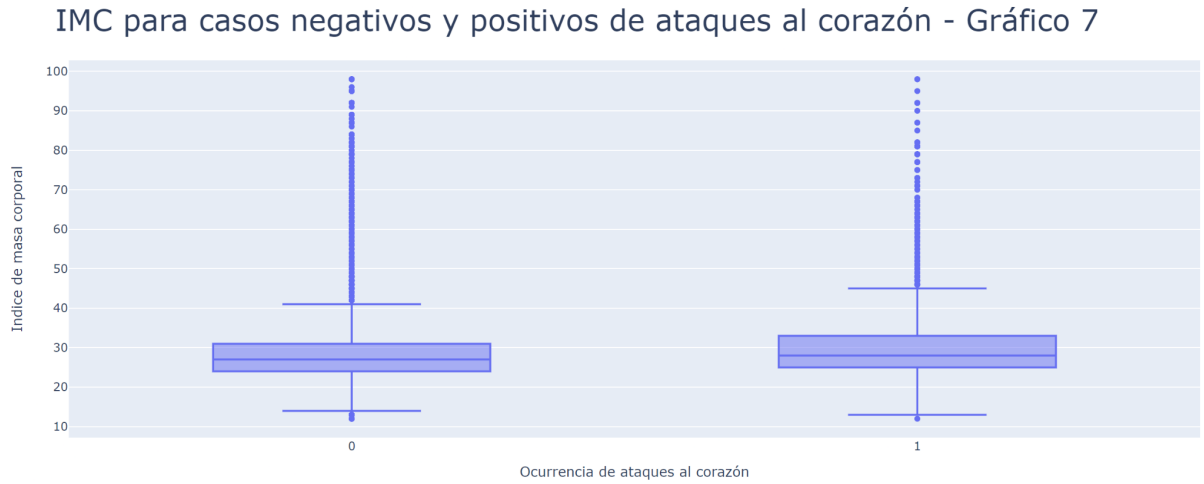


A partir del "Gráfico 6" se pudo observar que utilizando un umbral elevado (0.15) la variable target se encontró mayormente correlacionada con:

5. SaludGnrl (0.26)
6. DificCaminar (0.21)
7. Edad (0.22)
8. AltaPresionSanguinea (0.21)

Estos valores de correlaciones pertenecen al dataset original, el cual incluye targets positivos y negativos.

Gráfico 7



Al observar el "Gráfico 7" no fue posible encontrar algún tipo de patrón debido a la cantidad de datos extremos que se observaron. No se encontraron diferencias significativas en cuanto al IMC para los casos negativos y positivos de ataques al corazón.
El IMC no es significativo en la ocurrencia de enfermedades cardíacas (si se observa en el dataset original) respetando los insights obtenidos al observar el "Gráfico 5".

Selección de características

Se utilizó la selección secuencial hacia adelante para obtener las características con mayor relevancia. Dió como resultado las siguientes características:

a. AltaPresionSanguinea
b. AltoColesterol
c. IMC
d. Fumador
e. AccidenteCerebrovascular
f. Diabetes
g. SaludGnrl
h. SaludFisica
i. DificCaminar
j. Género
k. Edad

Luego se utilizó la eliminación bi-direccional para elegir las mejores características. Esta no dió buenos resultados dado que seleccionó demasiadas características que en pasos anteriores se vió que no tenían relevancia en nuestro análisis. Por lo que se descartó este método y se continuó trabajando con la selección secuencial hacia adelante.

Desbalance de dataset

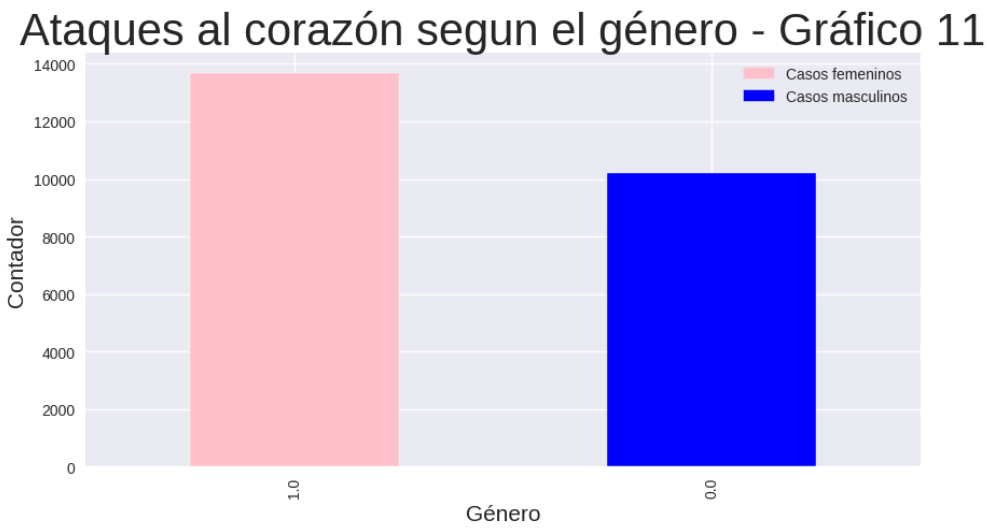
En el "Gráfico 8" se pudo observar un claro desbalance de clases.
Para resolver esto se utilizó una técnica de oversampling llamada SMOTE (Synthetic Minority Oversampling Technique). Esta es una técnica de sobremuestreo en la que se generan muestras sintéticas para la clase minoritaria, en este caso la clase que presentaba ausencia de enfermedades cardíacas. Este algoritmo ayuda a superar el problema de sobreajuste planteado por el sobremuestreo aleatorio. Se centra en el espacio de características para generar nuevas instancias con la ayuda de la interpolación entre las instancias positivas que se encuentran juntas.

Storytelling

Constantemente estamos presenciando noticias en donde se anuncian muertes rápidas y que parecen no tener ninguna causa. Según la Organización Mundial de la Salud (OMS), el accidente cerebrovascular es la 2ª causa principal de muerte a nivel mundial, responsable de aproximadamente el 11% del total de muertes. Más de 877,500 estadounidenses mueren de enfermedades cardíacas, accidentes cerebrovasculares y otras enfermedades cardiovasculares cada año. Quienes sufren de estas enfermedades pueden ser nuestras parejas, padres, abuelos, amigos o incluso conocidos. Se sabe que entre los diferentes géneros tenemos nuestras diferencias, ya sean físicas como psicológicas. Entonces, ¿se puede observar alguna diferencia en cuanto al género?

Para los siguientes gráficos se utilizó un dataset modificado que surgió de eliminar los casos negativos de enfermedades del corazón. Es decir que se utilizaron datos en donde todos los sujetos presentaban problemas cardiacos.

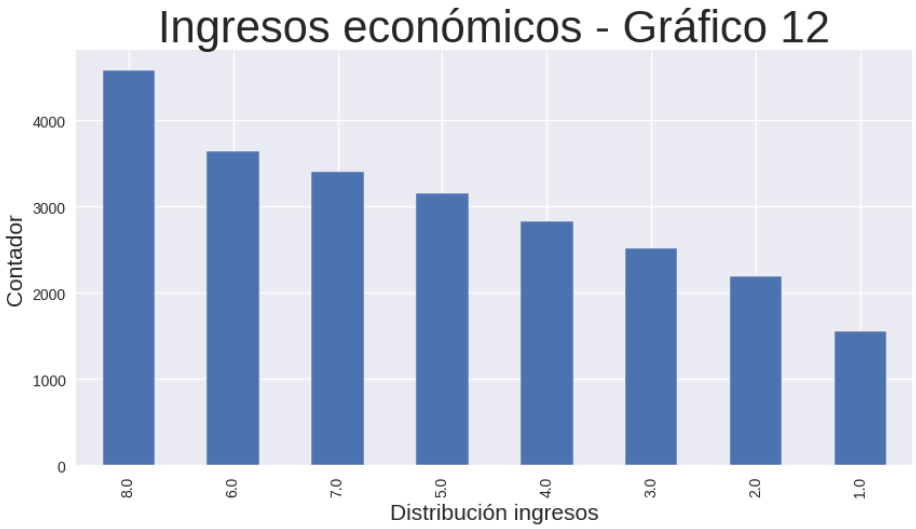
Gráfico 11



Como se pudo observar en el "Gráfico 11" aproximadamente el 58% de los casos es femenino y el 42% masculino. Por lo que no se pudo ver una diferencia clara del género en la presencia de problemas al corazón.

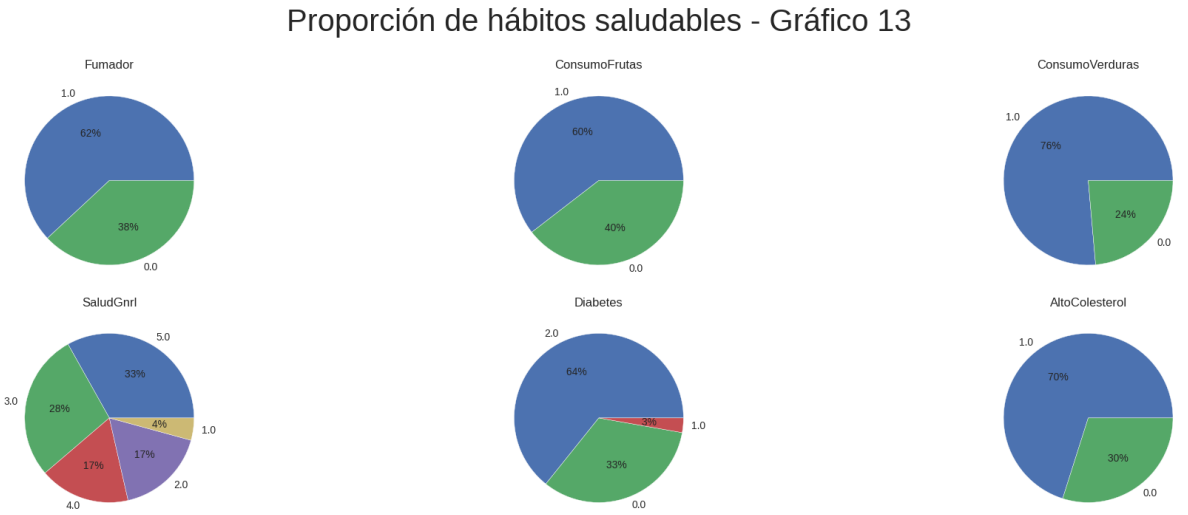
Entonces se presentó la pregunta, ¿las personas con mayor poder adquisitivo están exentos de los problemas al corazón?

Gráfico 12



Al observar el "Gráfico 12" se concluyó que seamos personas adineradas o no, nada nos exenta de sufrir problemas al corazón. Por otra parte, se debe tener en cuenta que aquellas personas con mayores ingresos tienen más acceso a instituciones de salud y son más propensas a testearse.
Otra pregunta que surgió fue: ¿Es posible identificar una mejora en las personas que presentan una buena alimentación y una buena condición física?

Gráfico 13



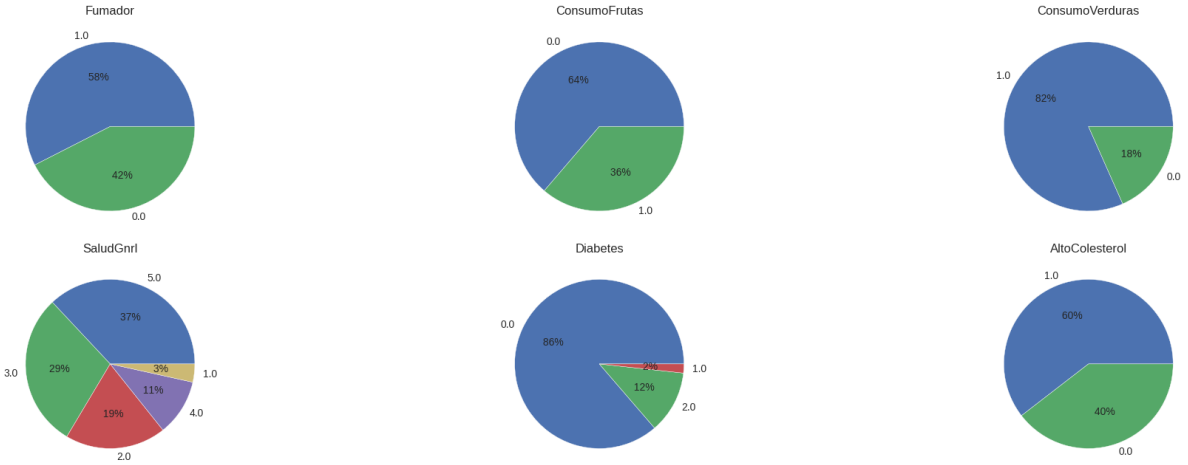
A primera vista se pudo concluir que la presencia de un buen estado de salud afecta a la presencia de ataques al corazón. Por otro lado, se debe saber que aquellas personas que sufrieron de algún ataque o han sido diagnosticadas con esta enfermedad comenzaron a cuidar su salud para evitar tener problemas mayores. Es por eso que se vieron estos datos.
En datos se observó que:

a. El 62% fueron fumadores
b. El 60% consumían frutas
c. El 76% consumían verduras

d. El 67% tenía diabetes
e. El 70% tenía alto colesterol

Gráfico 14 - Para este gráfico se utilizó una modificación del dataset original en donde solo se encontraban casos negativos de enfermedades cardíacas. Esto se dio con el objetivo de contrastar las diferencias entre casos positivos y negativos.

Proporción de hábitos saludables - Gráfico 14

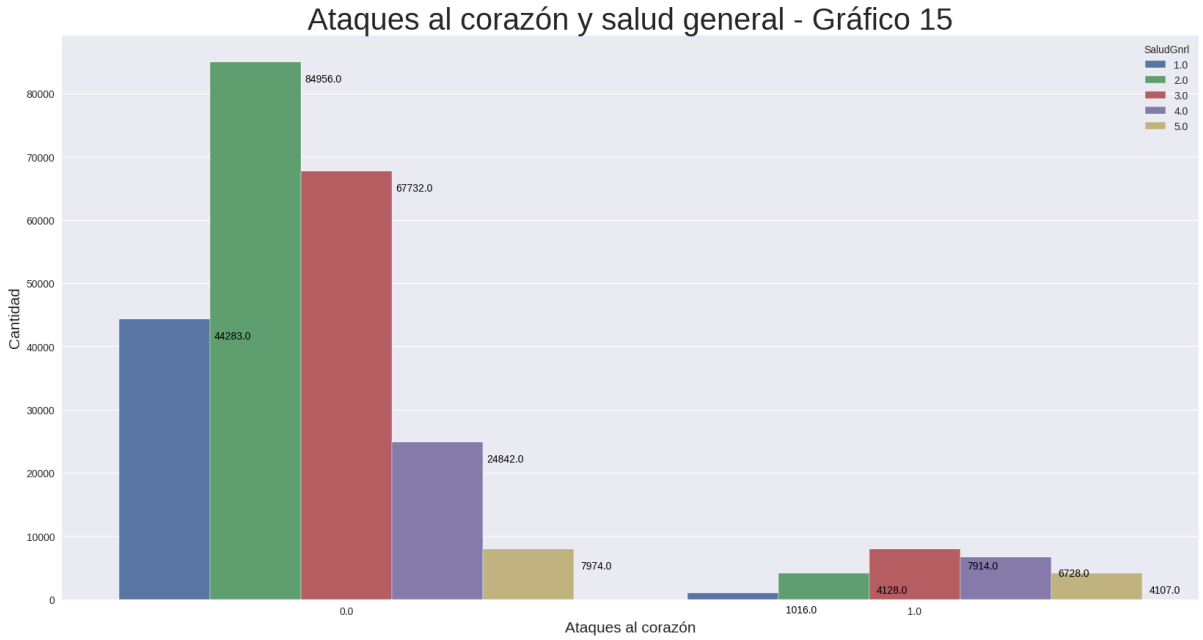


En datos se observó que:

a. El 58% fueron fumadores
b. El 64% consumían frutas
c. El 82% consumían verduras
d. El 88% tenía diabetes
e. El 60% tenía alto colesterol

Muchas estrategias de prevención de accidentes cerebrovasculares son las mismas que las estrategias de prevención de enfermedades cardíacas. Controlar la presión arterial alta, reducir la cantidad de colesterol y grasas saturadas en tu alimentación, hacer ejercicio de forma regular, evitar el alcohol y el tabaquismo entre otras recomendaciones para mantener un estilo de vida saludable. Es esencial la detección temprana de los distintos signos de advertencia para evitar o minimizar un accidente cerebrovascular. La solución está en nuestras manos y no se requiere de una inversión elevada más que el cuidado de uno mismo. El momento de actuar es hoy.

Gráfico 15



En el "Gráfico 15" se pudo observar que aquellas personas que:
No sufrieron de un ataque al corazón

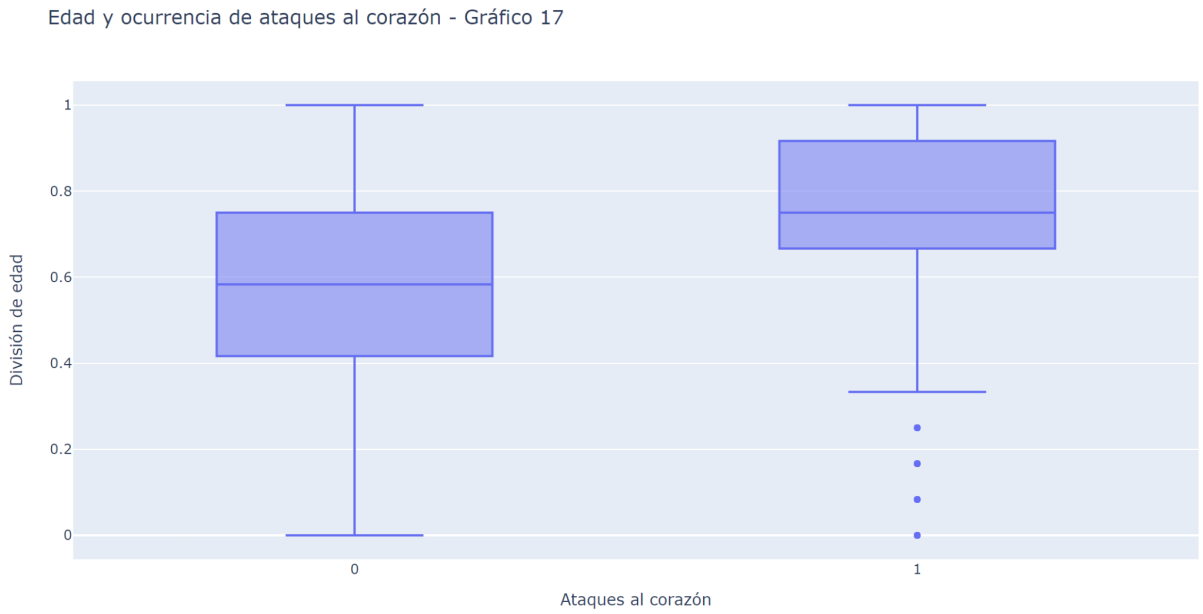
a. 85.7% tiene al menos un estado de salud general bueno

Sufrieron de un ataque al corazón

a. 54.6% tiene al menos un estado de salud general bueno

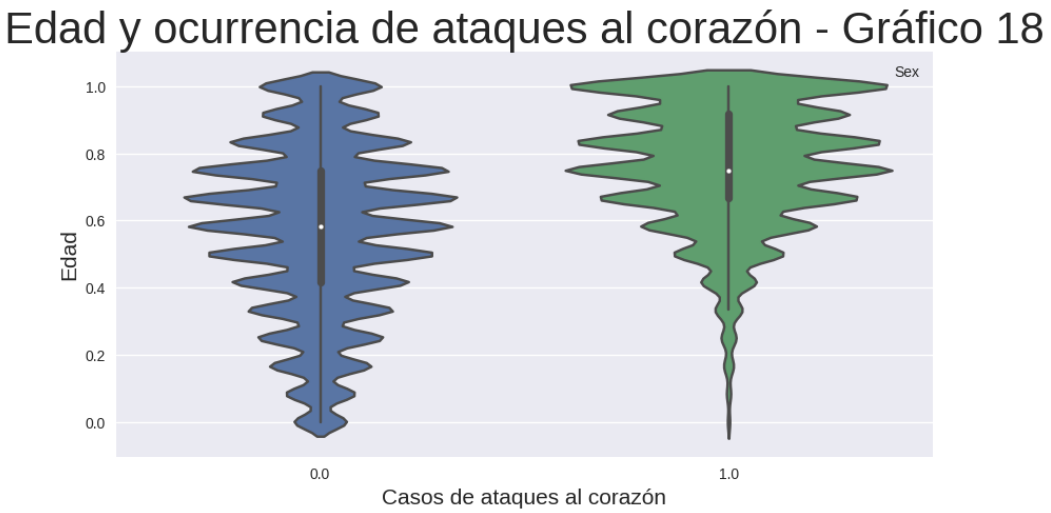
Claramente se puede ver un aumento del 30% del estado de salud general bueno en los casos en donde no surgieron de ningún ataque al corazón.

Gráfico 17



Al observar el "Gráfico 17" se pudo concluir que la mediana de los casos positivos fueron mayores a los correspondientes con los casos negativos. Es decir que aquellas personas de

edad avanzada tendieron a ser más susceptibles a sufrir ataques al corazón. Además se presentaron algunos casos extremos en los positivos, pero ninguno en los casos negativos.
Gráfico 18



En el "Gráfico 18" se pudo observar que los casos positivos de ataques al corazón se dieron en mayor medida en personas de edad avanzada.

Gráfico 19

En el "Gráfico 19" se pudo observar que aquellas personas que:
No sufrieron de un ataque al corazón

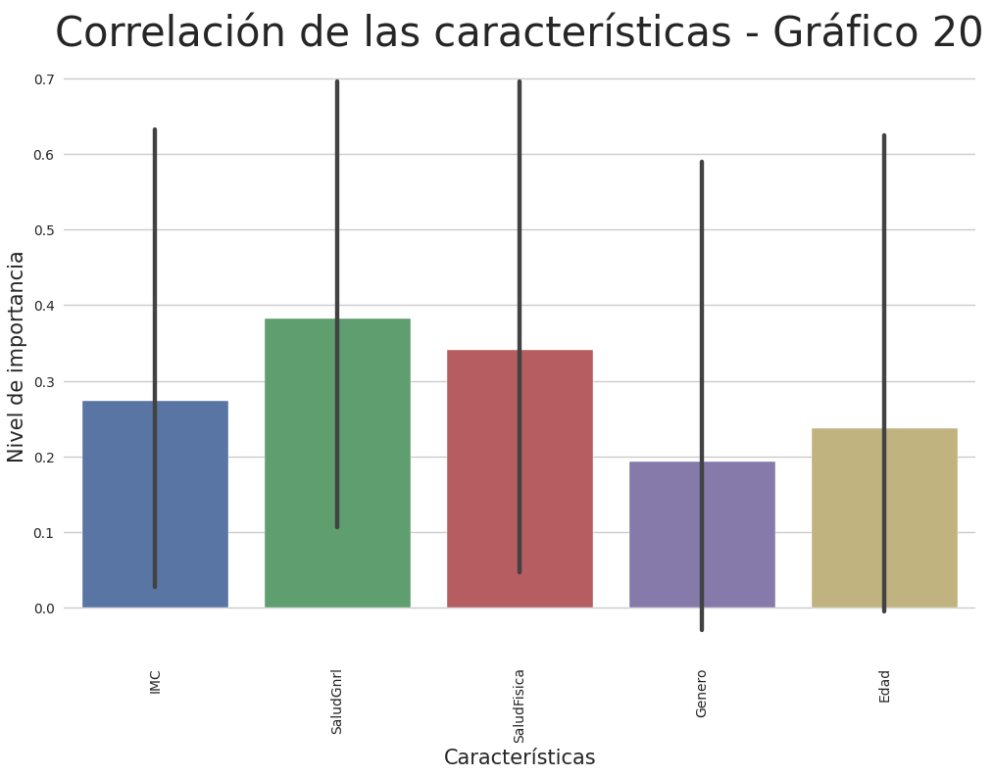
a. 97.2% no sufrió de un ACV
b. 2.8% sufrió de un ACV

Sufrieron de un ataque al corazón

a. 80.3% no sufrió de un ACV
b. 19.7% sufrió de un ACV

Claramente se puede ver un aumento de casi 10 veces más de casos de ACV en personas que sufrieron previamente un ataque al corazón.

Gráfico 20



En el "Gráfico 20" se pudo observar la correlación entre las variables más importantes para la determinación de nuestro target. Se pudo notar que todas superan el valor de 0.19 indicando una buena correlación entre las mismas.

Modelos de predicción

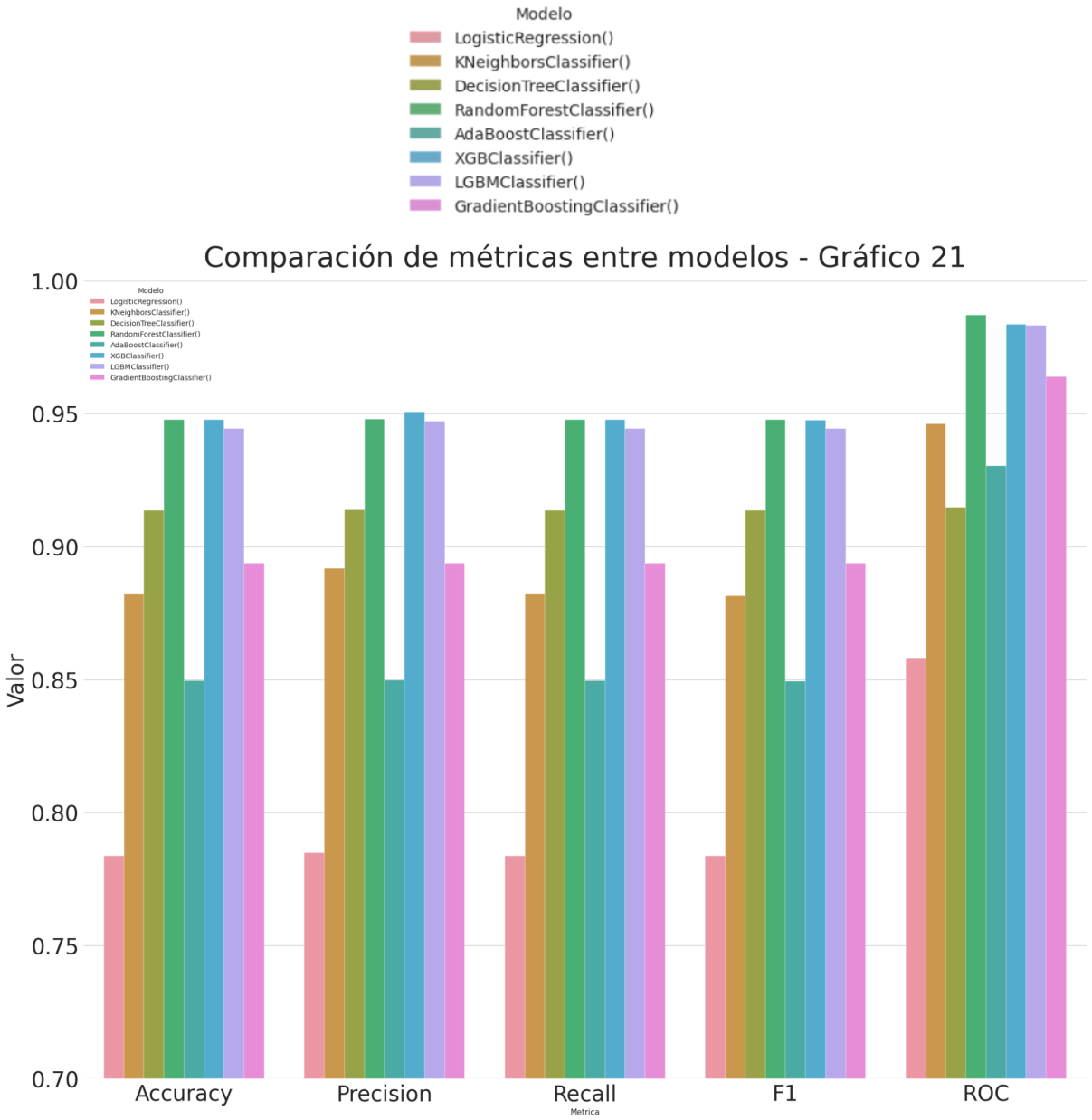
Se utilizaron distintos métodos de machine learning con el objetivo de predecir, dada ciertas características, si un sujeto tendrá enfermedades cardíacas o no. Se modeló utilizando:

a. LogisticRegression
b. KNeighborsClassifier
c. DecisionTreeClassifier
d. RandomForestClassifier
e. AdaBoostClassifier
f. XGBClassifier
g. LGBMClassifier
h. GradientBoostingClassifier

A continuación se observan algunos de las métricas obtenidas:

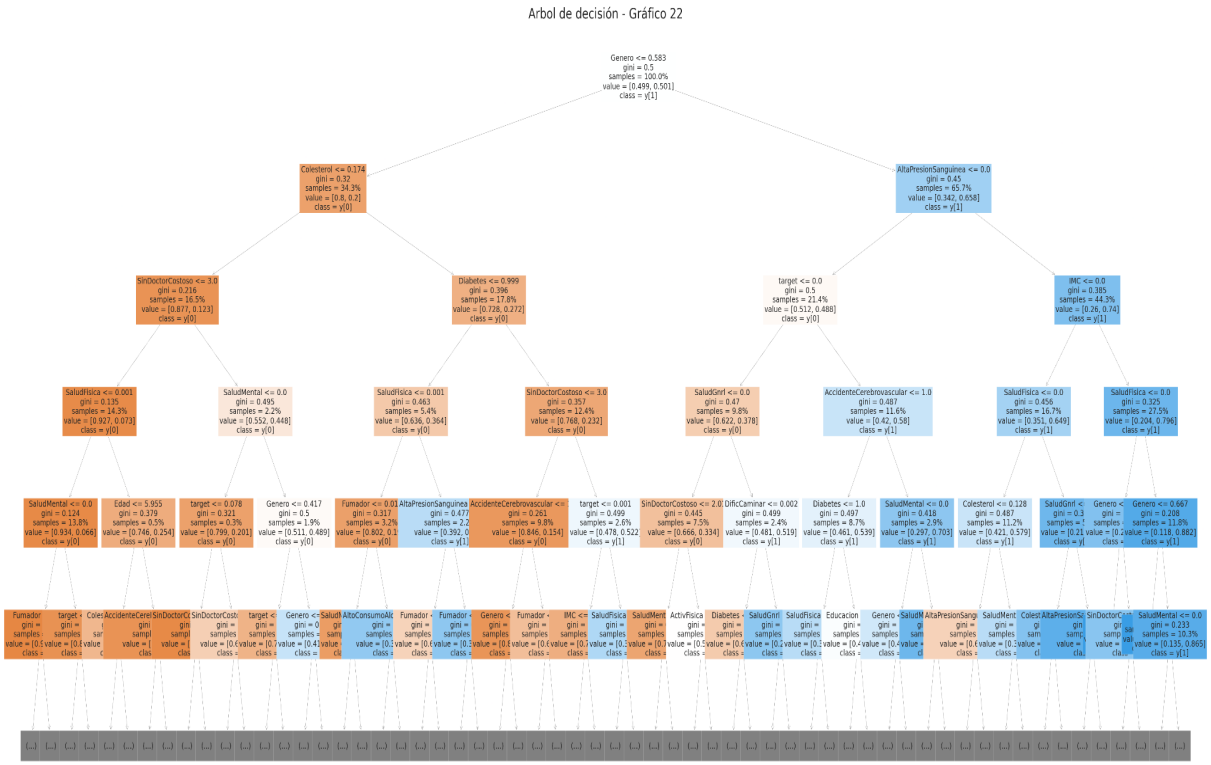
	Modelo	Accuracy	Precision	Recall	F1	ROC
0	LogisticRegression()	0.783759	0.784841	0.783759	0.783616	0.858134
1	KNeighborsClassifier()	0.882132	0.891810	0.882132	0.881474	0.946143
2	DecisionTreeClassifier()	0.913660	0.913749	0.913660	0.913660	0.914725
3	RandomForestClassifier()	0.947713	0.947875	0.947713	0.947705	0.987107
4	AdaBoostClassifier()	0.849471	0.849765	0.849471	0.849457	0.930437
5	XGBClassifier()	0.947691	0.950604	0.947691	0.947591	0.983625
6	LGBMClassifier()	0.944471	0.947114	0.944471	0.944373	0.983183
7	GradientBoostingClassifier()	0.893751	0.893755	0.893751	0.893749	0.964002

Para una mayor claridad se adjunta un gráfico comparativo que permite apreciar las métricas más importantes de estos modelos.



Se observó en el "Gráfico 21" que luego de modelar utilizando distintos métodos, la Clasificación por bosque aleatorio dió los mejores resultados. Teniendo en cuenta dos métricas fundamentales este modelo obtuvo valores muy buenos. Primero se tuvo en cuenta el accuracy que nos permitió saber la proporción de predicciones correctas del modelo, y luego el F1 el cual resulta importante dado que tiene en cuenta la precisión y el recall. Sin embargo este método obtuvo resultados similares al de los modelos de XGBClassifier y a LGBMClassifier() por lo que se podría haber optado por usarlos dada la ajustada diferencia. Por otra parte, el modelo de LogisticRegression dio los peores resultados.

Gráfico 22

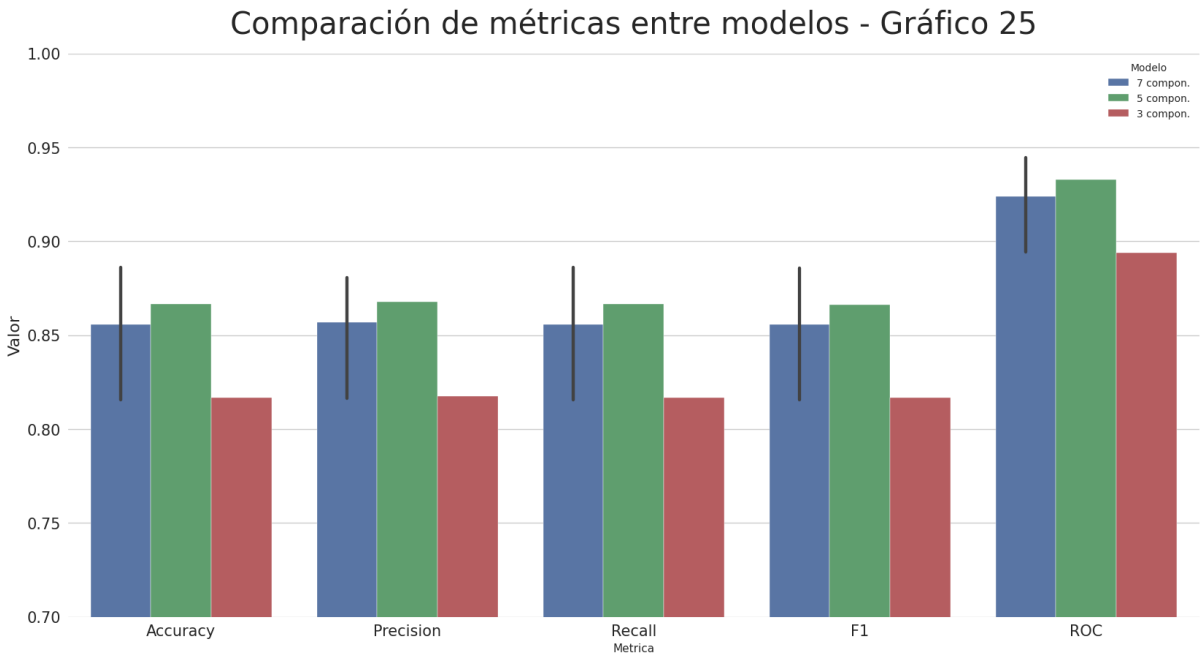


La visualización del "Gráfico 22" fue útil para comprender el funcionamiento interno del modelo y para obtener información sobre los datos. Los nodos superiores del árbol (Salud física, Accidente cerebrovascular, Sin dinero para un doctor) hicieron las divisiones más importantes. También contribuyó a tener una idea de qué características son más importantes para hacer predicciones.

Análisis de componentes principales

PCA es un método estadístico que permite simplificar la complejidad de espacios muestrales con muchas dimensiones a la vez que conserva su información. Fue utilizado para reducir la dimensionalidad del conjunto de datos, simplificar el análisis y mejorar el rendimiento. Se utilizó PCA con 7, 5 y 3 componentes. Luego se obtuvieron las métricas para cada ejecución.

Gráfico 25



En el gráfico 25 se pudo visualizar los valores de las distintas métricas para las distintas ejecuciones del algoritmo de PCA, entre las métricas se pudieron observar: Accuracy, precisión, recall, F1, ROC.
Se decidió utilizar la ejecución de PCA con 5 componentes la cual mostraba el mejor rendimiento en cuanto a poder de cómputo y métricas obtenidas. PCA con 5 componentes mostró:

1. Accuracy ≅ 0.88
2. Precisión ≅ 0.88
3. Recall ≅ 0.88
4. F1 ≅ 0.88
5. ROC ≅ 0.93

Conclusiones

Concluyendo con los datos obtenidos a lo largo de este trabajo se pudo determinar que las personas que sufrían de enfermedades cardíacas presentaron:

1. Alto colesterol (70%)
2. Alta presión en sangre (43%)
3. Diabetes (67%)
4. Alta ocurrencia del sexo femenino (58%)
5. Alto nivel de tabaquismo (62%)
6. Edad elevada (mayores de 64 años)
    a. La mitad de los sujetos que presentaban enfermedades cardíacas tenían por lo menos 75 años
    b. La mitad de los sujetos que no presentaban enfermedades cardíacas tenían por lo menos 58 años

    A partir de los 64 años, la cantidad de personas que sufrían de enfermedades cardíacas era mayor que aquellas que no tenían problemas.
7. Alto IMC (25)
    Sobre la cantidad de personas que presentaban un IMC mayor a 25 predominaban aquellas con enfermedades cardíacas.
8. Baja calidad en la alimentación
    Aquellas personas con enfermedades cardíacas presentaron:
        a. El 60% consumían frutas
        b. El 76% consumían verduras

    Por otra parte aquellos que no presentaron ninguna enfermedad:
        a. El 64% consumían frutas
        b. El 82% consumían verduras

El estado de salud general, los problemas para caminar, la edad y la alta presión sanguínea fueron las características que más afectan en la presencia de enfermedades cardíacas
Se concluyó que para prevenir enfermedades cardíacas se debe tener un control del colesterol, la diabetes, el tabaquismo y el índice de masa corporal. Esto se puede mejorar haciendo testeos anuales y mejorando la alimentación. Esto también está avalado con los datos obtenidos en el punto 7 los cuales mostraron una clara diferencia en el consumo de frutas y verduras. Otras caracteristicas que no pueden ser controladas como el sexo (femenino) y la edad (mayores de 64 años) tambien fueron determinantes al presentar enfermedades cardíacas.

Al analizar la presencia de accidentes cerebrovasculares se determinó que:
De las personas que no sufrieron de enfermedades cardíacas

a. 97.2% no sufrió de un ACV
b. 2.8% sufrió de un ACV
c. 85.7% tiene al menos un estado de salud general bueno

De las personas que no sufrieron de enfermedades cardíacas

a. 80.3% no sufrió de un ACV
b. 19.7% sufrió de un ACV

c. 54.6% tiene al menos un estado de salud general bueno

Se pudo ver un aumento de casi 10 veces más de casos de ACV en personas que sufrían previamente de enfermedades cardíacas.
Se pudo ver un aumento del 30% del estado de salud general bueno en los casos en donde no sufrían de enfermedades cardíacas.
Por lo que se pudo determinar que la presencia de enfermedades cardíacas afectaba en la aparición de un accidente cerebrovascular.

Limitaciones del trabajo

Por momentos se intentó personalizar los modelos de machine learning y deep learning. Dado al limitado uso de la versión gratuita de Google Colab no era posible terminar la ejecución y el uso del CPU no era el suficiente.
Al haber utilizado un dataset del año 2015 los datos pueden no tener validez en la actualidad. Una forma de resolver esto es aplicando un pipeline para automatizar el proceso de ETL. Haciendo uso de Buckets para almacenar la información, Airflow o Prefect para orquestar y establecer el tiempo de ejecución del pipeline y DBT para hacer transformaciones en los datos de forma automática podría ser una forma de mantener un dataset actualizado. De esta forma también sería posible hacer un análisis con una mayor cantidad de datos.