



Proyecto Final Data Science

Predicción de

ataques cardíacos

Integrantes:

- Carolina Abuelo
- Carolina Orse

30 de septiembre del 2022



Índice

1. Introducción

- a.* Problema de investigación
- b.* Objetivo principal
- c.* Hipótesis

2. Dataset

- a.* Descripción de las variables
- b.* Características principales

3. Análisis EDA

- a.* Estudio de las hipótesis
- b.* Matriz de correlación

4. Algoritmos

- a.* Transformaciones
- b.* Algoritmos de clasificación

5. Conclusión



Introducción


Las **enfermedades cardiovasculares** son aquellas que afectan al corazón y a todas las arterias del organismo. Su principal causa es la “aterosclerosis”, que es el depósito de placas de colesterol en el interior de las paredes de las arterias, provocando su obstrucción y comprometiendo la llegada de la sangre a órganos vitales como el corazón, el cerebro y el riñón. Por esta razón, la enfermedad arterial aterosclerótica es la principal causa del infarto agudo de miocardio (IAM), del accidente cerebrovascular (ACV) y de los aneurisma

Cuando se produce la obstrucción de una arteria y no hay flujo sanguíneo, se produce isquemia; si la arteria se rompe y hay pérdida de sangre, se produce hemorragia. La enfermedad coronaria es el compromiso aterosclerótico de las arterias que irrigan el corazón (las coronarias), pudiendo ocasionar un IAM.

A continuación detallamos los principales factores de riesgo cardiovascular:

- A mayor edad, mayor riesgo
- Antecedentes familiares de enfermedad cardiovascular prematura (padres/hermanos afectados antes de los 55 años en familiares hombres o antes de los 60 años en familiares mujeres)
- Tabaquismo
- Niveles elevados de colesterol en la sangre
- Presión arterial elevada (Hipertensión Arterial)
- Diabetes
- Sobrepeso y obesidad
- Inactividad física y estilo de vida sedentario
- Estrés crónico

En la actualidad, las enfermedades cardiovasculares (ECV) son las principales causas de muerte a nivel mundial. La Organización Mundial de la Salud (OMS) estimó en el 2015 que el 31% de todas las muertes mundiales fueron por ECV, siendo en total 17,7 millones de personas que murieron por ECV en los cuales el ataque cardíaco fue uno de los casos más comunes y potencialmente mortal (OMS, 2017).



La detección de ataques cardíacos es una tarea muy desafiante y más aún en países donde hay escasez de expertos y equipos humanos (Javeed & Zhou, 2016), los estudios revelan que en países de ingresos bajos y medios más del 80% de las defunciones son por esta causa y afectan casi por igual a hombres y mujeres (Kim & Kang, 2017). En lo general los expertos médicos llegan a diagnósticos basados en electrocardiogramas, ecografías, resultados de análisis de sangre y su experiencia personal, lo que aumenta los riesgos de errores y retrasa el tratamiento apropiado (OMS, 2017). Por lo que, en solución a estos problemas, investigadores han desarrollado diferentes sistemas inteligentes empleando diversos modelos, algoritmos y han usado como entrada las variables de los diagnósticos permitiendo tener mayor precisión para la detección automática de ataques cardíacos.

Problema y Objetivos de la Investigación

Objetivo principal

Construir un modelo de Machine Learning utilizando algoritmos de clasificación para que médicos, clínicos, aseguradoras y empresas den un diagnóstico más certero a la hora de prevenir un ataque cardíaco.

Problema de investigación

Identificar si existen patrones que puedan indicarnos que una persona sufrirá un ataque cardíaco

Hipótesis:

1. Los parámetros clínicos son predictores de mortalidad por ataque cardíaco
2. Las patologías clínicas (diabetes, anemia e hipertensión) son factores predictores de mortalidad por ataque cardíaco
3. El consumo de tabaco es un factor predictor de mortalidad por ataque cardíaco
4. La edad y el sexo son un factor predictor de mortalidad por ataque cardíaco

Dataset

Se utilizó un dataset del sitio web Kaggle cuyos autores son Davide Chicco, Giuseppe Jurman, quienes utilizaron la misma para su publicación “*Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone*” en la revista de investigación BMC Medical Informatics and Decision Making (2020).

Link Kaggle: <https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data>

Cuenta con un total de 299 observaciones y 13 variables numéricas.

Descripción de las variables:

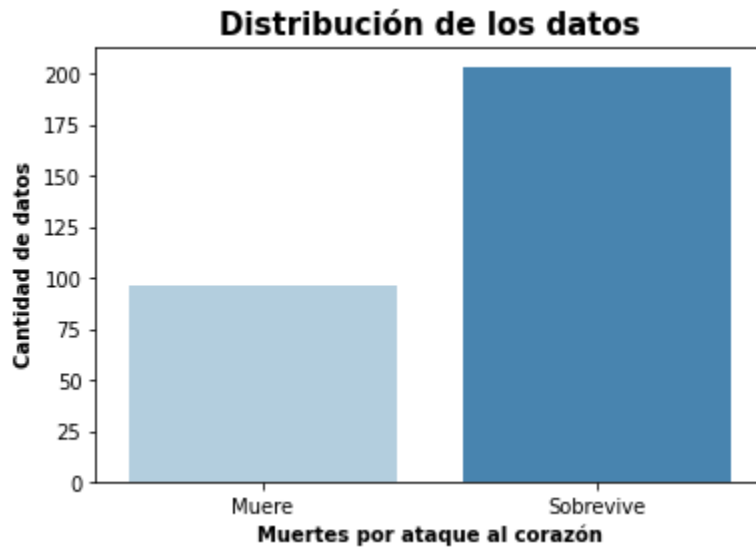
- ❖ Edad (float) - Edad del paciente
- ❖ Sexo (int) - Genero del paciente (Mujer=0, Hombre=1)
- ❖ Creatina Cinasa (int) - Nivel de la enzima CPK en sangre
- ❖ Fracción de eyección (int) - Medida del porcentaje de sangre que sale del corazón cada vez que este se contrae
- ❖ Plaquetas (float) - Cantidad de plaquetas en sangre
- ❖ Creatina en sangre (float) - Desecho generado por los músculos como parte de la actividad diaria
- ❖ Sodio en sangre (int) - Cantidad de sodio en sangre
- ❖ Diabetes (int) - Pacientes con la patología (Sano=0, Diabético=1)
- ❖ Anemia (int) - Pacientes con la patología (Sano=0, Anémico=1)
- ❖ Hipertensión (int) - Pacientes con la patología (Sano=0, Hipertenso=1)
- ❖ Fumador (int) - Consumo de tabaco del paciente (NO fumador=0, Fumador=1)
- ❖ Tiempo (int) - Tiempo que fue el paciente fue observado
- ❖ **Muerte (int) - Variable Target.** Indica si el paciente sufrió un ataque durante el periodo de observación (Sobrevive=0, Muere=1)

Características del Dataset

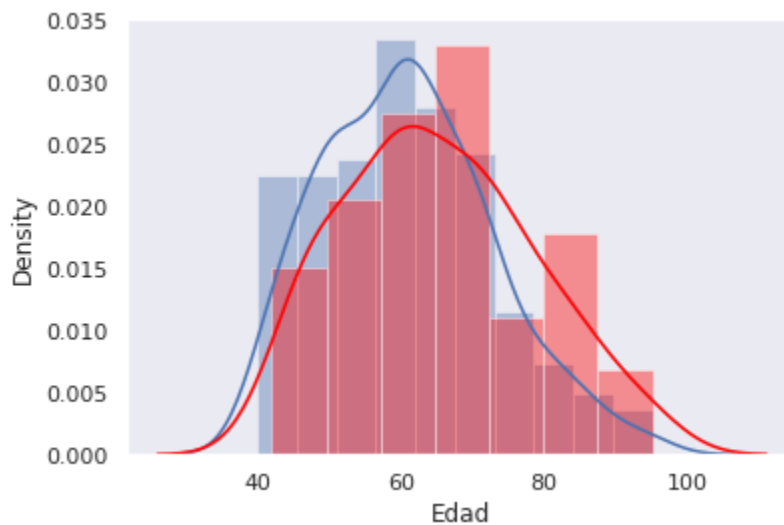
El mismo no poseía datos nulos como tampoco datos duplicados, es así que no requerimos utilizar ningún código para rellenar celdas vacías como tampoco eliminar datos duplicados, aunque estos últimos estimamos que tampoco hubieran inferido en el análisis realizado.

Comprendemos que al ser una observación en una población diversa, podrían haberse presentado casos similares y, sería sesgar la información, eliminar datos similares ya que pudieron haberse presentado en la muestra personas con características fisiológicas muy similares.

Nos encontramos que la distribución de los datos respecto a la variable target estaba balanceada, al contar con un 33,11% (99/299 casos) que murieron por ataque al corazón.



A su vez, visualizamos que la distribución entre los casos que sufren un infarto y la población en general respecto la edad son similares, determinando que la distribución de la variable target según la edad de los participantes de la muestra es normal.

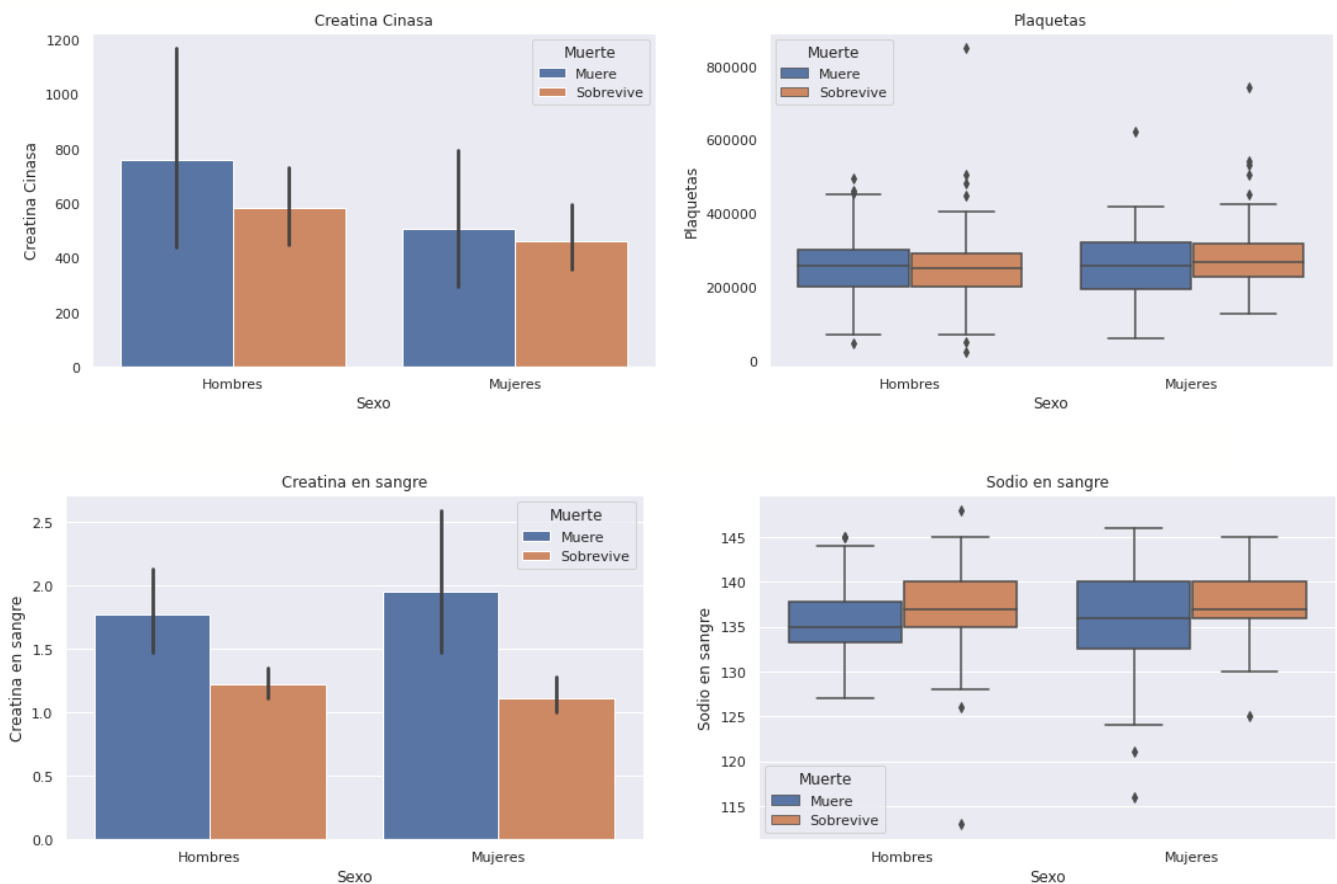


Análisis EDA (Exploratory Data Analysis)

Análisis de las hipótesis

Con el fin de responder las hipótesis previamente planteadas, analizamos las distintas variables que el dataset nos proporciona.

H1: Los parámetros clínicos son predictores de mortalidad por ataque cardíaco.

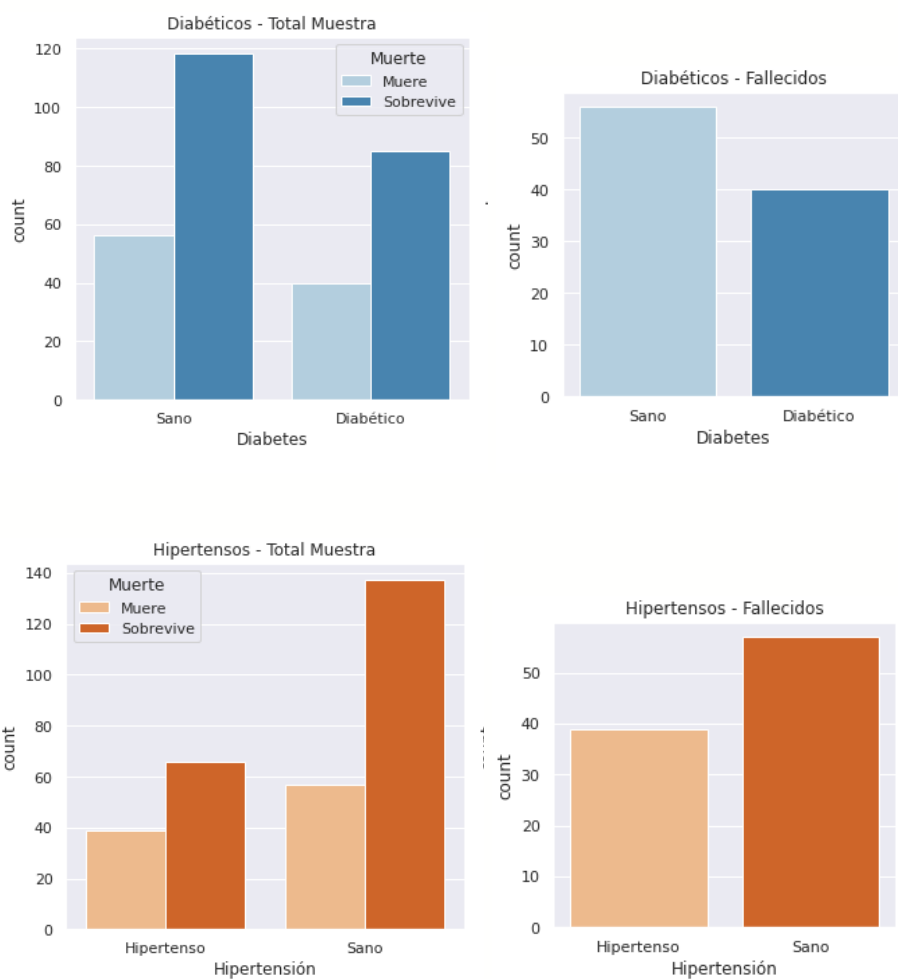


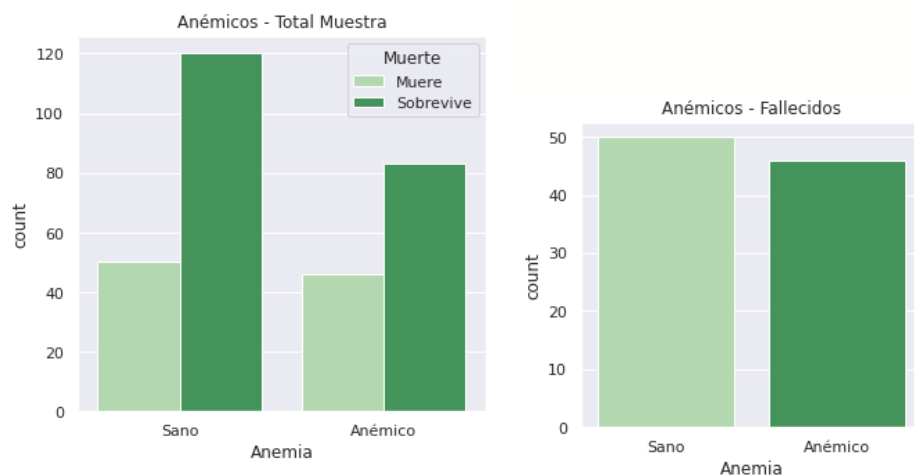
En base a lo observado en los gráficos, podemos deducir que aquellos que fallecieron por ataques cardíacos poseen valores más altos de creatina cinasa como creatina en sangre, afectando esta última mayormente a la población femenina que tiene un porcentaje aún mayor.

Respecto al sodio en sangre, se presenta un fenómeno de hiponatremia ya que los valores de aquellos que fallecen por ataque son inferiores a los parámetros normales.

Por último, las plaquetas en ambos géneros no presentan variación frente a la variable target.

H2: Las patologías clínicas (diabetes, anemia e hipertensión) son factores predictores de mortalidad por ataque cardíaco.





En los primeros gráficos visualizamos que si el paciente cuenta con una única enfermedad preexistente (diabetes, anemia o hipertensión) ninguna por sí sola es determinante para dar lugar a un ataque cardíaco.

En un análisis más detallado de la muestra de fallecidos, armamos el siguiente ranking de patologías que generan mayor predisposición a tener un ataque cardíaco:

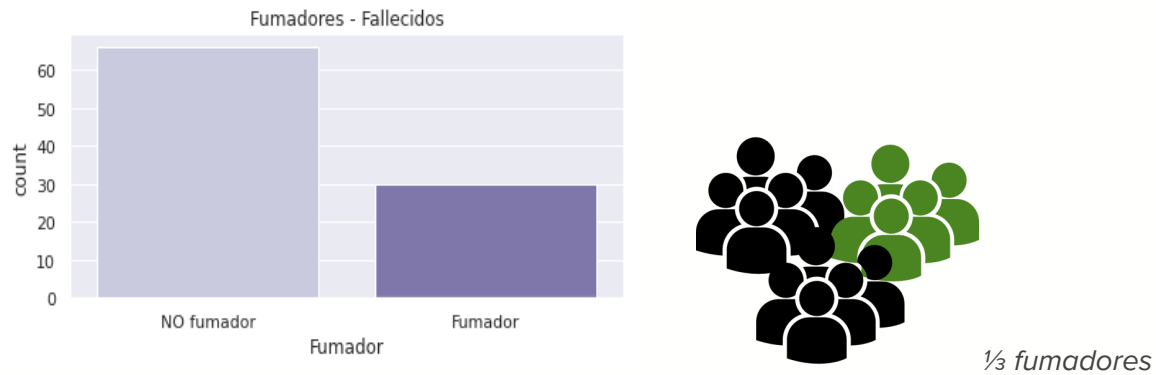
1° Anemia: de la muestra analizada, es decir de aquellos casos que fallecieron por un ataque cardíaco, un 45% de la muestra eran anémicos.

2° Diabetes: un 40% de los observados poseían dicha patología.

3° Hipertensión: con una leve diferencia del segundo puesto, un 39% de la muestra eran hipertensos.

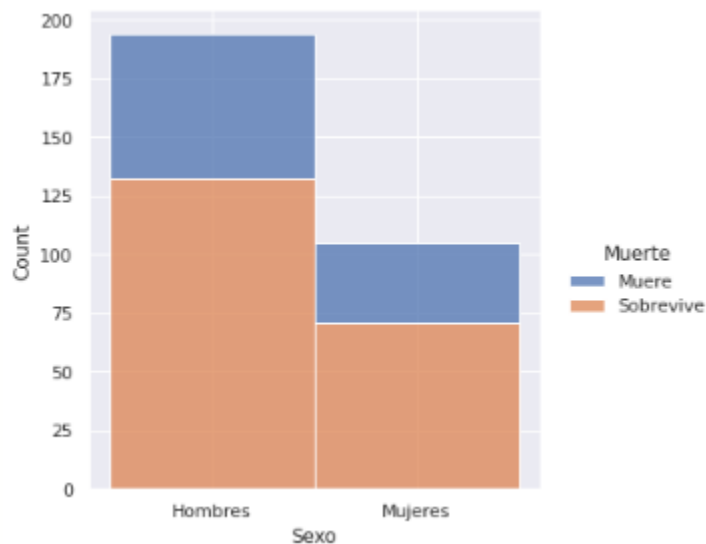
Queda pendiente en un próximo estudio, analizar si las comorbilidades (la suma de más de una patología) da lugar a aumentar aún más el riesgo de sufrir dicho ataque.

H3: El consumo de tabaco es un factor predictor de mortalidad por ataque cardíaco



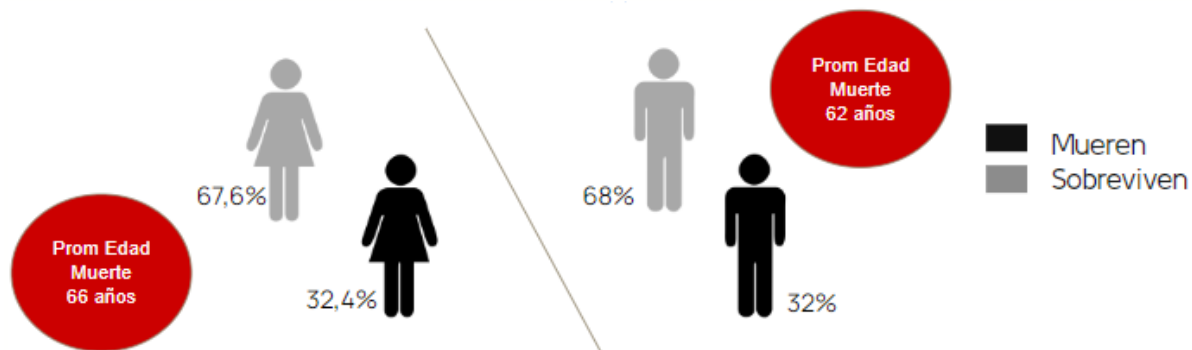
Respecto a los hábitos de consumo, podemos inferir que el consumo de tabaco no es un factor determinante a la hora de predecir pacientes con posibles ataques cardíacos, ya que de los pacientes que fallecieron por enfermedades cardiovasculares sólo un tercio eran fumadores.

H4: La edad y el sexo son un factor predictor de mortalidad por ataque cardíaco



Mediante los gráficos inferimos que los hombres son aquellos más propensos a sufrir un infarto pero, haciendo un análisis a mayor detalle, la muestra está en términos nominales conformada

en mayor parte por hombres, siendo esto un sesgo en la muestra.



Analizando cada género por separado, visualizamos que en ambos casos el porcentaje de fallecidos es del 32%, mientras que el porcentaje de sobrevivientes es del 68%. Es así que no podemos deducir que algún sexo tenga una mayor probabilidad y predisposición genética a fallecer por ataques cardíacos.

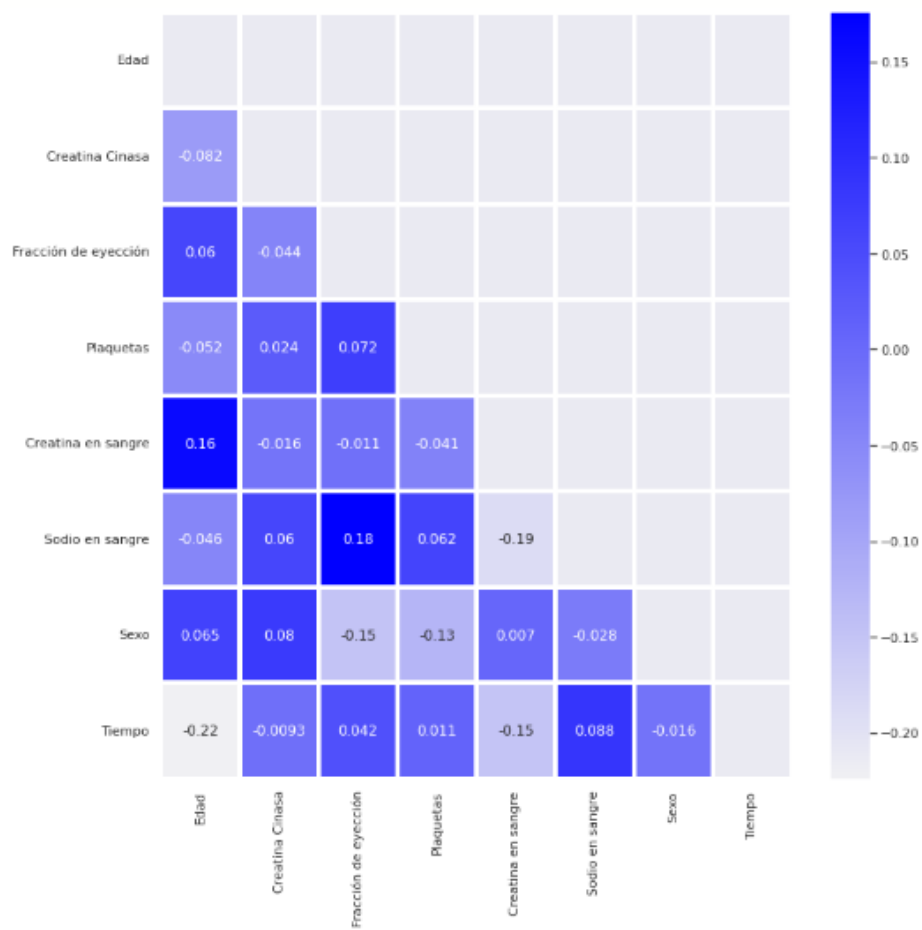
Dentro de dicho análisis, también pudimos determinar que el promedio de edad de muerte en las mujeres es de 66 años, mientras que el de los hombres es de 62 años.

Análisis de la correlación de las variables

Matriz de correlación

La matriz de correlación muestra los valores de correlación de Pearson, que miden el grado de relación lineal entre cada par de elementos o variables.

En nuestra muestra, la matriz de correlación indica que la creatina cinasa y el tiempo son las variables con mayor correlación. A su vez, la variable que más se relaciona con la posibilidad de tener un ataque cardíaco es la creatina cinasa.



Algoritmos

Transformaciones sobre el dataset

De todas las variables que nos brindaba el dataset, seleccionamos algunas de las variables en base a su importancia previo a correr los modelos. Entre las variables que suprimimos, se encuentra el tiempo y la fracción de eyección.

A la hora de determinar la importancia de las variables, la matriz de correlación indicaba que el tiempo tenía una alta influencia en el resto de las variables pero, dado que el dataset no nos brindaba información precisa respecto al tiempo que estaba siendo medido en la muestra y, por esto mismo, cualquier inferencia que hagamos estaría sesgada plenamente a la subjetividad de nuestra interpretación, decidimos eliminar la misma.

Respecto la fracción de eyección, no era una variable de alta relevancia como tampoco se nos brindaba información precisa de su influencia en nuestra observación, siendo así que al presentar la misma valores outliers, tomamos la decisión de también de eliminarla.

Algoritmos de clasificación

Éstos algoritmos resuelven problemas de clasificación que necesitan predecir la clase más probable de un elemento, en función de un conjunto de variables de entrada. Para este tipo de algoritmos, la variable target o respuesta, es una variable de tipo categórica.

En base al problema de investigación descrito al principio, en el cual buscamos identificar si existen patrones que nos indiquen que una persona padecerá un ataque cardíaco, decidimos analizar los modelos:

- **Regresión Logística**
- **Random Forest Classifier**
- **Árbol de Decisión**

Random Forest, al igual que el Árbol de decisión, son modelos de aprendizaje supervisado comúnmente utilizados en problemas clasificación, aunque también puede usarse para

problemas de regresión.

Por otro lado, la Regresión Logística es una técnica de aprendizaje automático que proviene del campo de la estadística. No es un algoritmo, sino que es un método para problemas de clasificación.

Tras definir los parámetros y realizar el entrenamiento de dichos modelos, obtuvimos las siguientes métricas:

	Precision	Recall	Accuracy
Regresión logística	0.61	1.00	0.63
Random Forest Classifier	0.65	0.91	0.66
Árbol de Decisión	0.64	0.77	0.61

Frente a los resultados obtenidos, se llevaron a cabo mejoras mediante las siguientes herramientas de optimización:

- RandomizedSearchCV
- Cross Validation.

Tras este proceso, todos los modelos tuvieron una notoria mejora en su performance, arrojando los siguientes resultados:

	Accuracy		Accuracy
Regression logistica	63%		72%
Random Forest Classifier	66%	→	76%
Arbol de Decision	61%		74%

El modelo seleccionado por tener la mejor performance fue **Random Forest Classifier**. Dicho modelo arroja el mejor resultado de la métrica seleccionada (Accuracy) tanto antes como después de ser optimizado, obteniendo 66,7% previa a la optimización y un 76,13% tras las mejoras.

En su versión "Classifier", es de los mejores modelos a utilizar en problemas de clasificación, permitiendo reducir la dimensionalidad del dataset y brindando un mejor rendimiento de generalización durante el entrenamiento. Dicha generalización se logra al compensar los errores de las predicciones de los distintos árboles de decisión.



Conclusión

Tras el análisis de distintos algoritmos de clasificación, el modelo Random Forest Classifier superó las expectativas que se tenían con respecto a su performance. Tras las optimizaciones, la tendencia de dicho modelo se mantiene, aunque los otros modelos también hayan presentado una mejora en sus métricas.

A los efectos del estudio realizado, consideramos que hubiera sido interesante contar con más variables de tipo social (entiéndase estado civil, profesión, grupo familiar), geográficas (lugar de residencia, nacionalidad), como también mayor información clínica de cada paciente respecto a otros factores de riesgo (antecedentes familiares de enfermedades cardíacas, peso, ibm, patologías psiquiátricas, etc.). A su vez, de haber contado con un número superior de casos, hubiéramos obtenido hallazgos más significativos para el problema planteado.

Finalmente, pensando en un modelo de negocio, los próximos pasos consistirán en buscar inversores interesados en la problemática planteada, que financien tanto la investigación como el desarrollo del proyecto. Es así, que inicialmente requerimos desarrollar una aplicación frontend que nos permita interactuar con el modelo para luego presentar la misma a los distintos inversores, entre ellos: empresas de salud, aseguradoras, hospitales y entidades gubernamentales del ámbito de salud y otros.

