

**Trabajo Final DataScience II**:

**Modelo de predicción de infarto agudo de miocardio.**

**Alumno**: Luciano Battioni

Data science 2024 comisión 60940

última versión:

**Docente**: Jorge Ruiz

**Tutor**: Fabricio di Palma

Tabla de contenido

[Problemática a resolver 3](#_Toc176814500)

[Modelado 3](#_Toc176814501)

[Adquisición de datos 4](#_Toc176814502)

[Diccionario de variables. 4](#_Toc176814503)

[Descripción del data set 5](#_Toc176814504)

[Data *Wrangling* y *EDA* 7](#_Toc176814505)

[Transformación de variables: 7](#_Toc176814506)

[Análisis de distribución y *outliers* 8](#_Toc176814507)

[Análisis bi y tri variado. 9](#_Toc176814508)

[Variables de interés: 10](#_Toc176814509)

[Feature Engineering 10](#_Toc176814510)

[Modelo de Regresión logística 11](#_Toc176814511)

[Modelo de árbol de decisión 11](#_Toc176814512)

[Modelo XGBOOST 11](#_Toc176814513)

Introducción

Las enfermedades cardiovasculares son la principal causa de muerte a nivel mundial. Dentro de estas la enfermedad coronaria (EC) es la más frecuente. Su importancia global ha llevado a investigadores del siglo XX a desarrollar numerosos modelos predictores de eventos cardiovasculares (infarto agudo de miocardio, accidente cerebrovascular, muerte de causa vascular), tanto para la población general como para poblaciones particularmente vulnerables (pacientes diabéticos, pacientes con enfermedad renal crónica, etc.) con la esperanza de poder intervenir terapéuticamente y poder prevenirlos. Sin embargo, su validez externa y generalizabilidad han demostrado tener importantes limitaciones. Esto se ha debido, entre otras causas, a que se derivaron de cohortes de bajo número de pacientes, poco representativas de determinadas etnias o se construyeron pensando en la factibilidad de ser usados en consultorio por lo que solo utilizan variables seleccionadas, comprometiendo, tal vez, parte su precisión.

Las técnicas de *Machine learning* (ML) procesan gran cantidad de variables sin necesidad de intervención de un clínico. Esto haría factible la implementación de modelos más precisos, desplegables en ámbitos como EHR (*Electronic Health Records*) que permitan la estimación de riesgo cardiovascular en base a la información proporcionada por el sistema de manera automática.

# Problemática a resolver

Mediante los datos clínicos y demográficos se intentará realizar un modelo que permita predecir el si una persona tendrá un evento coronario (EC)

Se minimizará la tasa de falsos negativos y se maximizará la sensibilidad ya que esta herramienta se utilizará como *screening* poblacional. De esta manera se identificarán paciente de alto riesgo y se realizarán intervenciones de prevención de bajo costo y alto impacto como modificación en los hábitos higiénico dietéticos, manejo del stress, cesación tabáquica, etc.

Sensibilidad= TP / (FN+TP)

VPN= VN/ (VN+FN)

# Modelado

## Adquisición de datos

Los datos fueron obtenidos de una base disponible en la página [Kaggle](https://www.kaggle.com/datasets/iamsouravbanerjee/heart-attack-prediction-dataset.). En una primera etapa se generó un diccionario de variables. De esta manera se estudió que datos contienen y que información clínica representan.

## Diccionario de variables.

1. **Patient ID**: texto, identificador del caso

2. **Age**: numérica edad del paciente

3. **Sex**: dicotómica (Male/Female)

4. **Cholesterol**: numérica, nivel de colesterol en mg/dl

5. **Blood** **Pressure**: texto, contiene tanto diastólica como sistólica separadas por barra invertida.

6. **Heart** **Rate**: numérica, frecuencia cardíaca

7. **Diabetes**: dicotómica, si el paciente tiene diabetes

8. **Family** **History**: dicotómica, si tiene antecedentes familiares de enfermedad cardíaca (1: Yes, 0: No)

9. **Smoking**: dicotómica, si el paciente fuma actualmente (1: Smoker, 0: Non-smoker)

10. **Obesity**: dicotómica, si el paciente es obeso (1: Obese, 0: Not obese)

11. **Alcohol** **Consumption**: categórica, el nivel de consume de alcohol.(None/Yes)

12. **Exercise** **Hours** **Per** **Week**: numérica, número de horas de actividad física por semana

13. **Diet**: categórica, hábitos dietarios(Healthy/Average/Unhealthy)

14. **Previous** **Heart** **Problems**: dicotómica, antecedentes de problemas cardíacos del paciente (1: Yes, 0: No)

15. **Medication** Use: dicotómica, si el paciente usa medicación (1: Yes, 0: No)

16. **Stress** **Level**: categórica, nivel de stress del paciente(1-10)

17. **Sedentary** **Hours** **Per** **Day**: numérica, horas de actividad Sedentaria por día

18. **Income**: numérica, ingresos anuales en USD

17. **BMI**: numérica, el resultado de dividir peso/altura\*\*2

19. **Triglycerides**: numérica, nivel de triglicéridos en mg/dl

20. **Physical** **Activity** **Days** **Per** **Week**: numérica, días de actividad física por semana

21. **Sleep** **Hours** Per Day: numérica, número de horas de sueño por día

22. **Country**: nominal, procedencia del paciente

23. **Continent**: nominal, continente del país de procedencia

24. **Hemisphere**: nominal, Hemisferio del continente del país de procedencia

25. ***Heart******Attack******Risk***: dicotómica, presencia de riesgo de infarto (1: Yes, 0: No). Variable target

# Descripción del data set

El dataset contiene 8763 filas y 26 columnas, de las cuales 1 de ellas es la variable target: *heart\_attack\_risk*. Las demás variables se utilizarán como variables independientes en la medida que sumen valor predictor al modelo. Para ello que analizará la importancia de cada *feature* en un modelo inicial y se re entrenará un segundo modelo utilizando la menor cantidad de variables. De esta manera se optimizarán procesos de cómputo.

El dataset no posee filas repetidas ni datos faltantes. El dataset, es ficticio, por ende, tiene distribución homogénea de todas sus variables. Finalmente, no posee *outliers* ni valores erróneos.

De las 26 columnas, 2 son variables tipo *string*: patient\_id y blood\_pressure. Esta última debido a que tiene tanto valores de tensión arterial sistólica y diastólica separados por “/”. Esta variable será recodificada en la sección *data* *engineering*. Diez variables son numéricas: age, colesterol, heart\_rate, exercise\_hours\_per\_week, sedentary\_hours\_per\_day, income, BMI, triglycerides, sleep\_hours\_per\_day y physical\_activity\_days\_per\_week, sin embargo, esta será interpretada como categórica debido a su naturaleza discreta y ordinal (no puede hacerse 1.5 días de ejercicio a la semana, y siempre es mejor hacer 6 días de ejercicio por semana que 3). El resto de las variables son categóricas. Nominales: sex, diabetes, family\_history, smoking, obesity, previous\_heart\_problems, medication\_use, country, continent, hemisphere, heart\_attack\_risk. Ordinales: alcohol\_consumption, diet,stress\_level

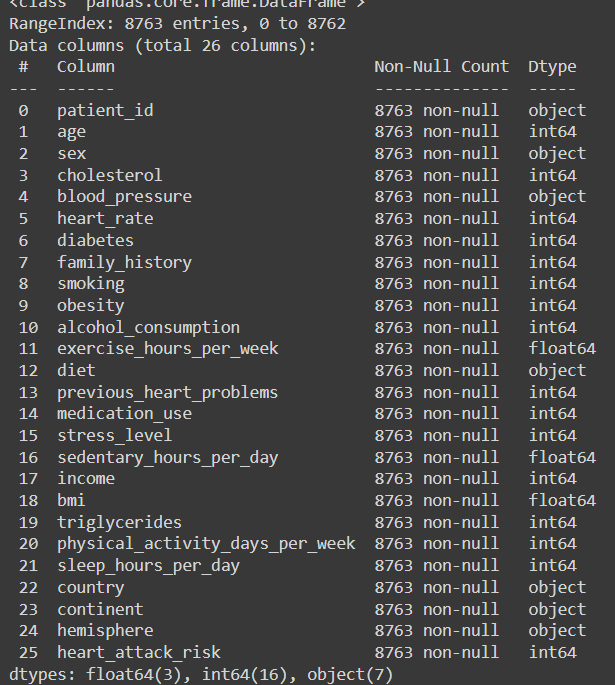


Fig. 1: Data *types* e *info* del Data set sin transformar

Objetivos

Para el primer entregable trabajaremos e intentaremos comprobar las siguientes hipótesis:

1. Relacionar los factores de riesgo cardiovascular con la cantidad de hs de ejercicio y sedentarismo. Entendiendo que es más probable que los pacientes con mayores factores de riesgo son los más sedentarios.

2. Relacionar los niveles de tensión arterial sistólica y diastólica con los niveles de Stress

3. Relacionar país, región con la dieta. Entendiendo que la idiosincrasia local afecta hábitos alimenticios.

Estas hipótesis serán contestadas con técnicas de EDA y análisis estadístico clásico

4 La hipótesis para el trabajo final es que se puede generar un modelo de ML que permita predecir el evento de infarto agudo de miocardio.

# Data *Wrangling* y *EDA*

## Transformación de variables:

Variables numéricas fueron llevadas al tipo que menor memoria requiere.

La variable blood pressure que se transformó en dos variables numéricas eliminándose el “/” y la columna original posteriormente.

*df\_raw['bps']=[pd.to\_numeric(i.split('/')[0],downcast='integer') for i in df\_raw.blood\_pressure]*

*df\_raw['bpd']=[pd.to\_numeric(i.split('/')[1],downcast='integer') for i in df\_raw.blood\_pressure]*

*df\_raw=df\_raw.drop(columns='blood\_pressure')*

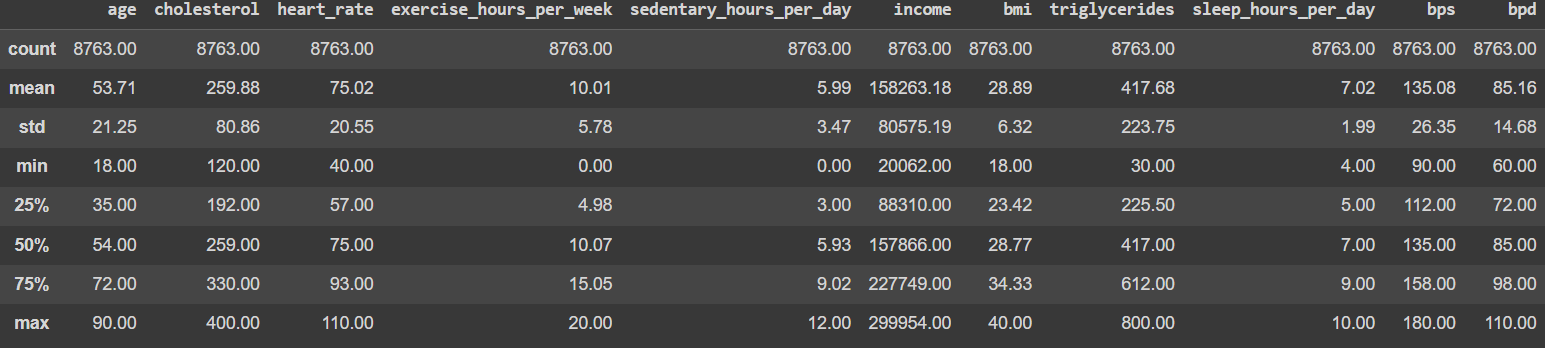
Las variables tipo *object* como país, continente, etc. que no tienen un orden de jerarquía se pasaron el tipo categoría.

Las variables que originalmente estaban catalogadas como numéricas por estar ya en codeadas se pasaron a categoría y se les asignó una jerarquía cuando correspondiera.

*df\_raw.diet=df\_raw.diet.cat.reorder\_categories(['Healthy', 'Average', 'Unhealthy'], ordered=True)*

*df\_raw.stress\_level=df\_raw.stress\_level.cat.reorder\_categories([1,2,3,4,5,6,7,8,9,10], ordered=True)*

*df\_raw.physical\_activity\_days\_per\_week=df\_raw.physical\_activity\_days\_per\_week.cat.reorder\_categories([0,1,2,3,4,5,6,7], ordered=True)*

Fig 2. Análisis descriptivo univariado de variables cuantitativas

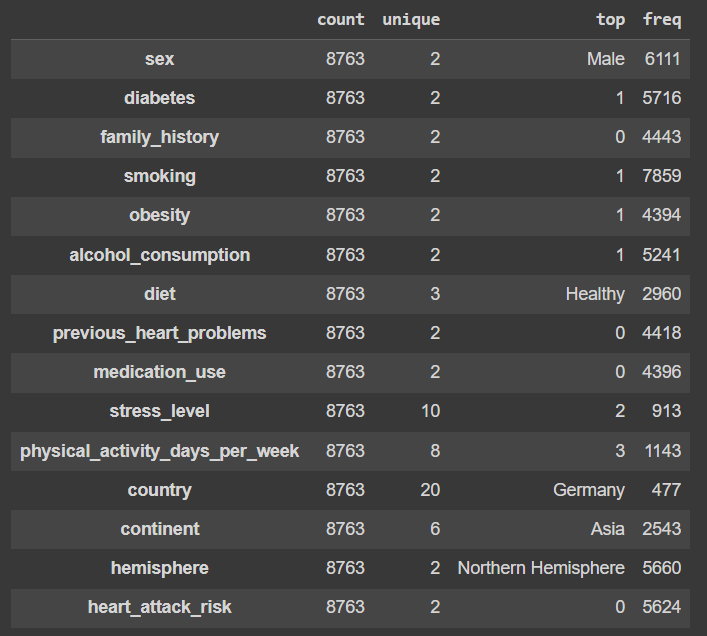


Fig 3: análisis de las variables categóricas y object

## Análisis de distribución y *outliers*

La demostración de este análisis debido al tamaño del data set no se reproducirá en este documento. Para esto se remite al lector al Notebook acompañante. Se puede decir que las distribuciones de todas las variables son no normal. Y si bien algunas variables presentan outliers se encuentran dentro del espectro de posibilidades biológicas, por lo que no se imputarán ni se eliminarán.

Objetivos 1 al 3

# Análisis bi y tri variado.

Los gráficos de este análisis no serán reproducidos en este documento, se remite al lector al notebook acompañante.

**Objetivo 1:**

Para este análisis se generaron 3 objetos:

*frcv\_cat=['sex','diabetes','family\_history','smoking','obesity','diet','previous\_heart\_problems','stress\_level']*

*frcv\_con=['age','cholesterol','bmi','triglycerides']*

*target=['exercise\_hours\_per\_week','sedentary\_hours\_per\_day','physical\_activity\_days\_per\_week']*

Con ellos se graficaron las relaciones y se testearon las hipótesis nulas entre las variables independientes y dependientes.

Los resultados de estos análisis fueron los siguientes:

La mediana de *'exercise\_hours\_per\_week'* en las categorías de *frcv\_cat* no presentó diferencias significativas utilizando wilcoxon rank sum test

La mediana de *'sedentary\_hours\_per\_day'* en las categorías de *frcv\_cat* no presentó diferencias significativas wilcoxon rank sum test

El test de chi cuadrado con corrección para múltiples testeos de Bonferroni no mostró diferencias significativas entre ‘*stress\_level' y 'physical\_activity\_days\_per\_week'*

Se utilizó el test de Spearman para testear el nivel de correlación entre *frcv\_con* y *target.* Se observó una correlación positiva débil 0.02 p= 0.046 para *'cholesterol'* y *'exercise\_hours\_per\_week'*

**Objetivo 2:**

Mediante el test de Kruskal-Wallis se compararon las medias de la tensión arterial sistólica y diastólica a través de las categorías de niveles de stress. No se observaron diferencias significativas.

**Objetivo 3:**

Para testear esta hipótesis se generó un objeto que contiene las tres variables de interés: *“country”, “continent” y “hemisphere”.*

No se observaron variaciones en la dieta a lo largo de los países, continentes o hemisferios.

Objetivo 4

# Variables de interés:

La variable target: *Heart\_attack\_risk*

Ya que esta es una variable binaria y desbalanceada se requieren algoritmos de clasificación binaria y pre procesamiento que tenga en cuenta su frecuencia despareja.

# Feature Engineering

Como se dijo previamente no hay valores perdidos. Se quitó la columna de patient ID ya que no es necesaria. Las variables categóricas se las transformó mediante *Labelencoders*

La matriz de correlación no muestra relaciones fuertes entre las variables. Hay una relación moderada entre tabaquismo y edad, así como también hay correlaciones moderadas entre las variables de país, continente y hemisferio, pero esto último era de esperarse.

Primer análisis de modelos de clasificación

Se corrieron lo siguientes modelos de clasificación:

Árbol de decisión

RandomForest

XGboost

Regresión logística binaria.

Se dividió el dataset en 80% entrenamiento y validación y el 20% restante en testeo.

Se implementaron técnicas de over sampling y under sampling debido a que es un dataset desbalanceado (SMOTETomek), así como también se realizó una Standard scaler para poder realizar la regresión logística.

El resultado fue el siguiente:

Las sensibilidades y VPN de los modelos en la validación cruzada fueron de mejor a peor:

-Árbol de decisión, Sens: 0.74, VPN 0.63

- XGboost, Sens: 0.85, VPN 0.65

- Regresión logística, Sens 0.45, VPN 0.62

RandomForest, Sens 0.99, VPN 0.75 (que se no se testeó más a fondo debido a que se contaba con XGBOOST y se sospechó que podía ya estar realizando sobreajuste, como se verá luego con los otros modelos basados en árboles)

# Modelo de Regresión logística

En el primer entrenamiento se determinaron los coeficientes absolutos de los features del dataset. Se tomaron los primeros 15 features más relevantes y se buscaron los mejores hiperparámetros con un HalvingridsearchCV.

El resultado fue un modelo que tiene una sensibilidad del 58% y un VPN de 57% en los datos de entrenamiento + validación.

En el set de testeo, disminuye la sensibilidad y mejora el VPN.

# Modelo de árbol de decisión

Se sigue el mismo procedimiento descripto en la regresión logística. Obteniéndose en el set de entrenamiento y validación una sensibilidad de 80% y un VPN 77%.

Sin embargo, la sensibilidad del set de testeo cae al 51% con un VPN de 65%.

Al intentar buscar un punto de corte óptimo a través del método de la curva ROC, se observa que la probabilidad óptima es cercana al 50%.

El gráfico del árbol de decisión es muy complejo para visualizarse adecuadamente.

# Modelo XGBOOST

El modelo XGBOOST hizo un overfit en la corte de entrenamiento y validación, clasificando positivamente a todos los casos lo mismo que en el set de testeo.

Luego de corroborar que no hubo cambios con la implementación de un punto de corte óptimo, se procedió a realizar un nuevo halving grid search ajustando el rango de hiperparámetros para disminuir el overfitting.

Sin embargo, el resultado fue el mismo, es decir la clasificación como positivos todos los casos.

Conclusión final.

Se puede concluir que no se pudo construir una herramienta de screening con los datos obtenidos de manera sintética.