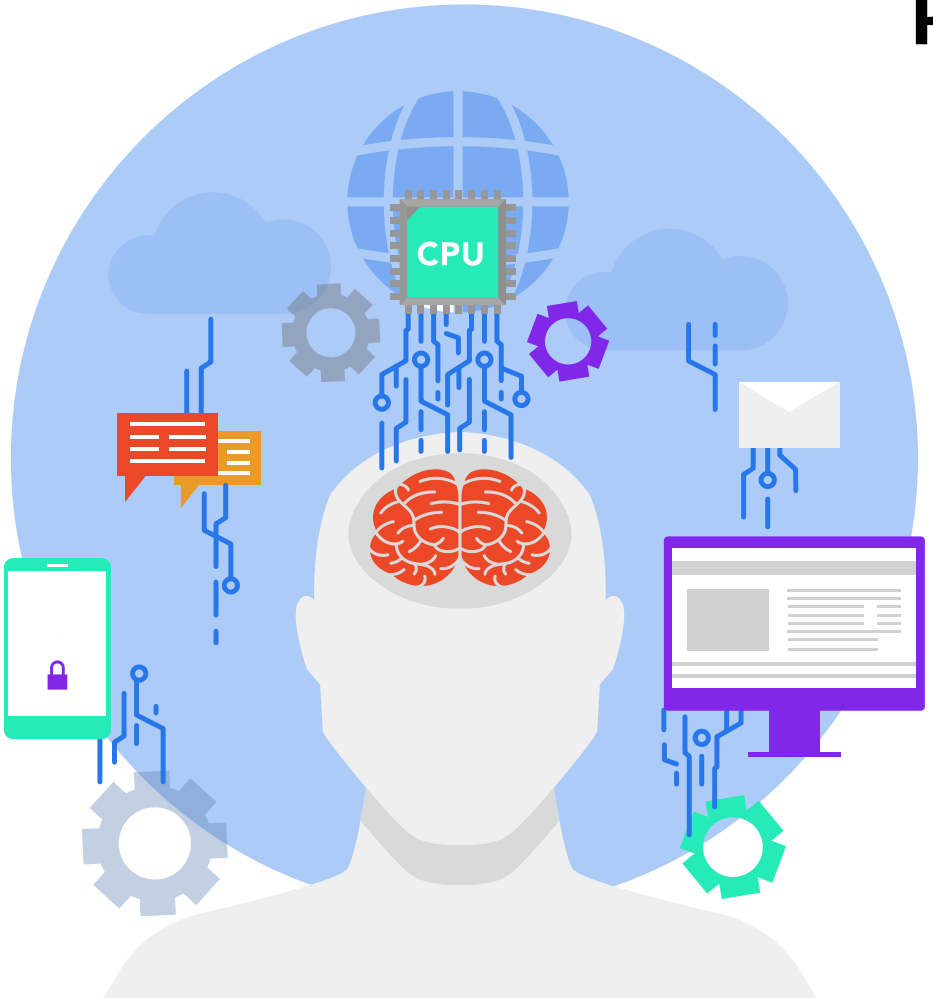
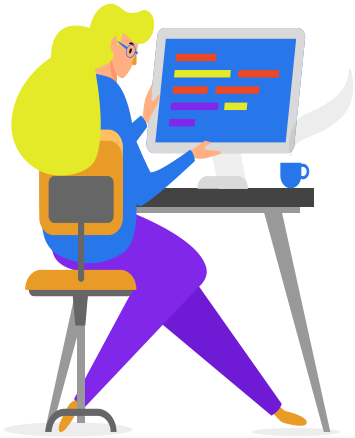


Predicción de Enfermedades Cardíacas a través del Análisis de Hábitos de Vida

DATA SCIENCE II - CODER

MaríaLaura Zulatto - 2024



Indice

MOTIVACIÓN Y AUDIENCIA

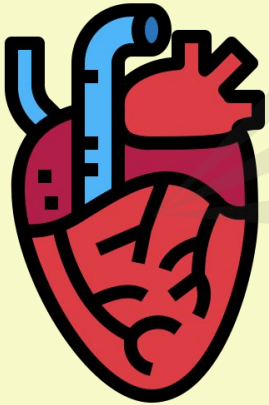
MOTIVACIÓN

- **Problema Global:** Enfermedades cardíacas como principal causa de muerte.
- **Complejidad de los Factores:** Interacción de múltiples hábitos y factores.
- **Innovación en Prevención:** Uso de aprendizaje automático para identificar patrones clave.

AUDIENCIA

- **Profesionales de la Salud**
- **Instituciones de Salud Pública**
- **Investigadores en Aprendizaje Automático**
- **Público General**
- **Aficionados del sistema cardiovascular (Como yo)**

HIPÓTESIS Y PREGUNTAS DE INTERÉS



¿Qué papel juegan los factores demográficos (edad, sexo, etc) en la incidencia de enfermedades cardíacas?

¿Cómo se encuentran los niveles de colesterol y triglicéridos de la población?

¿Cuál es el modelo de aprendizaje que mejor se ajusta a nuestro conjunto de datos?

METADATA Y DATA WRANGLING.

DATASETS

Fuente 1 : Kaggle
Muestras:8763
Variables: 26

Dataset extremadamente balanceado, no útil para análisis completo por sí solo.

Fuente 2 : Kaggle
Muestras:4239
Variables: 16

Complemento para las variables de interés de nuestro análisis.

Input data

METADATA:
12947 rows × 16 columns

Numérica continua

- Age
- Heart Rate
- Systolic
- Diastolic
- Cholesterol
- Triglycerides
- Glucose
- BMI
- Exercise Hours Per Week
- Sedentary Hours Per Week
- Prevalent Cardiovascular

Binaria

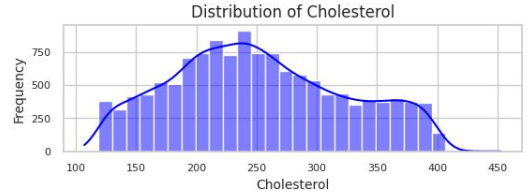
- Diabetes
- Medical Use
- Smoking
- Sex

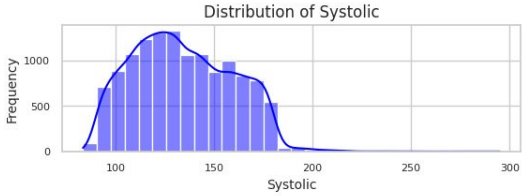
Output

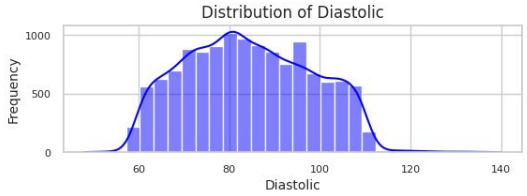
Heart Attack Risk
1: Yes
0: No

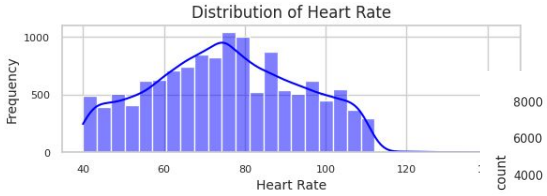
EDA : Distribuciones de variables

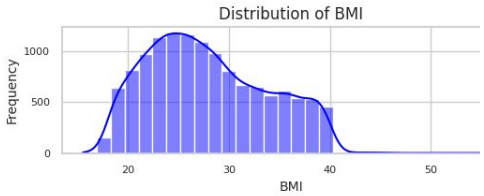
(posterior a limpieza e imputación de nulos)

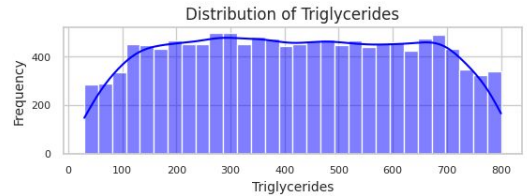


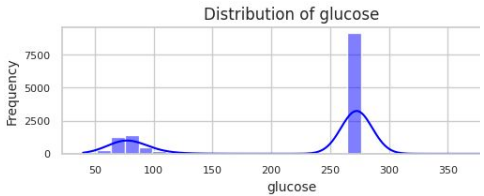


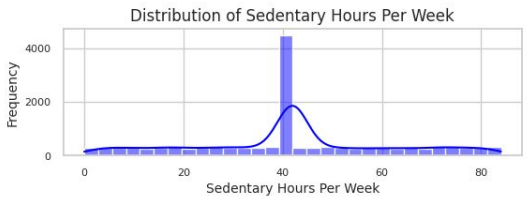


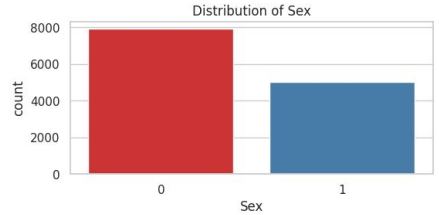


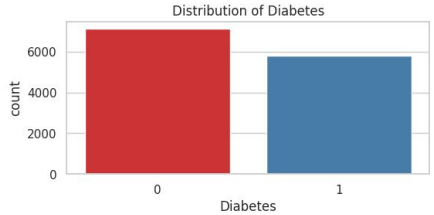


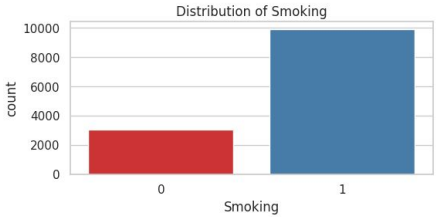


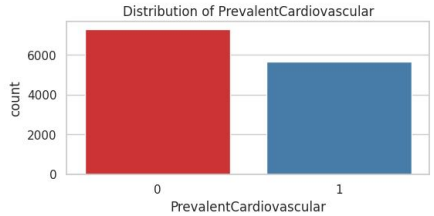


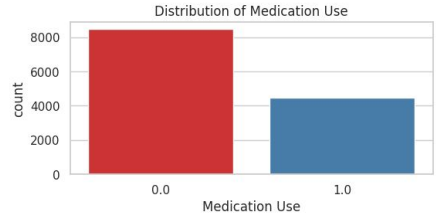


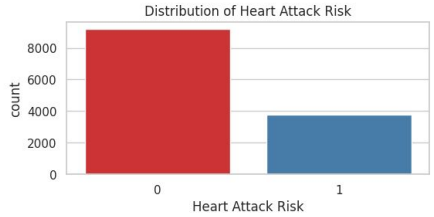




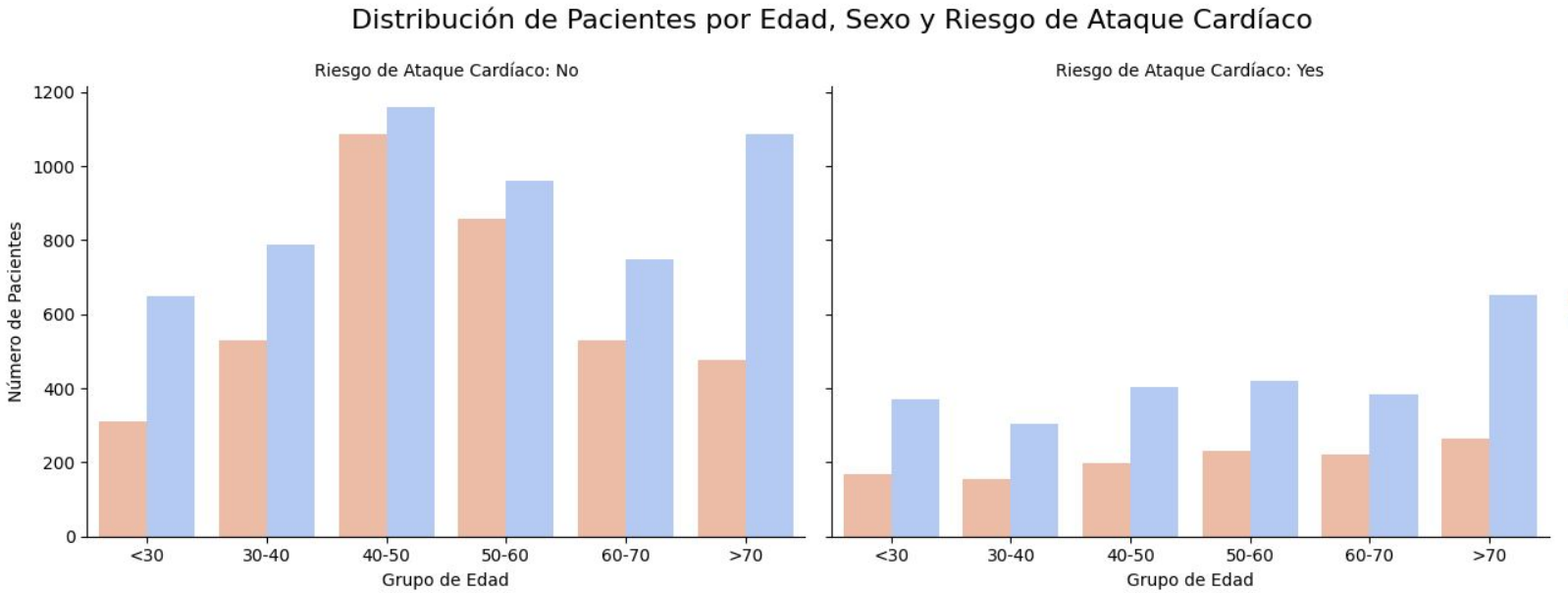








Análisis bivariado : Factores Demográficos en enfermedades cardíacas



7922

5025

Insight:

- Los hombres en el conjunto de datos tienen una mayor prevalencia de factores de riesgo asociados con enfermedades cardíacas
- mayor número de pacientes en los rangos de edad de 40-70 años podría correlacionarse con un aumento en el riesgo de ataque cardíaco con la edad.

Análisis bivariado : Niveles de colesterol y triglicéridos.

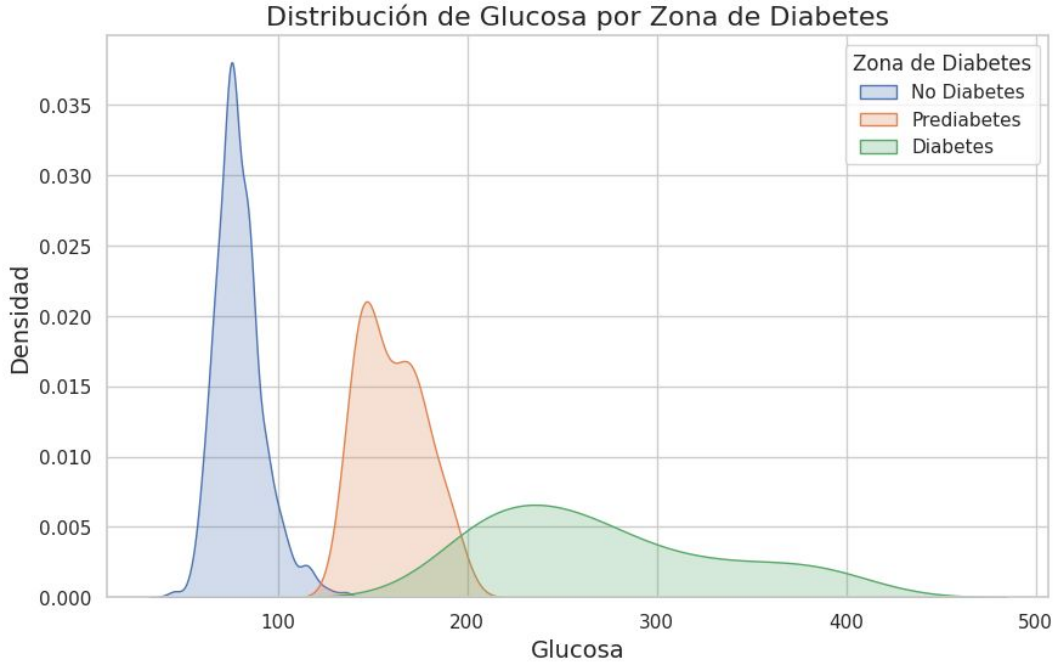
Insight : La población maneja niveles de triglicéridos y colesterol por arriba de los límites normales.

Implicancia directa a formación de placas de ateroma

Se imputaron los triglicéridos con una distribución uniforme, manteniendo su distribución original.

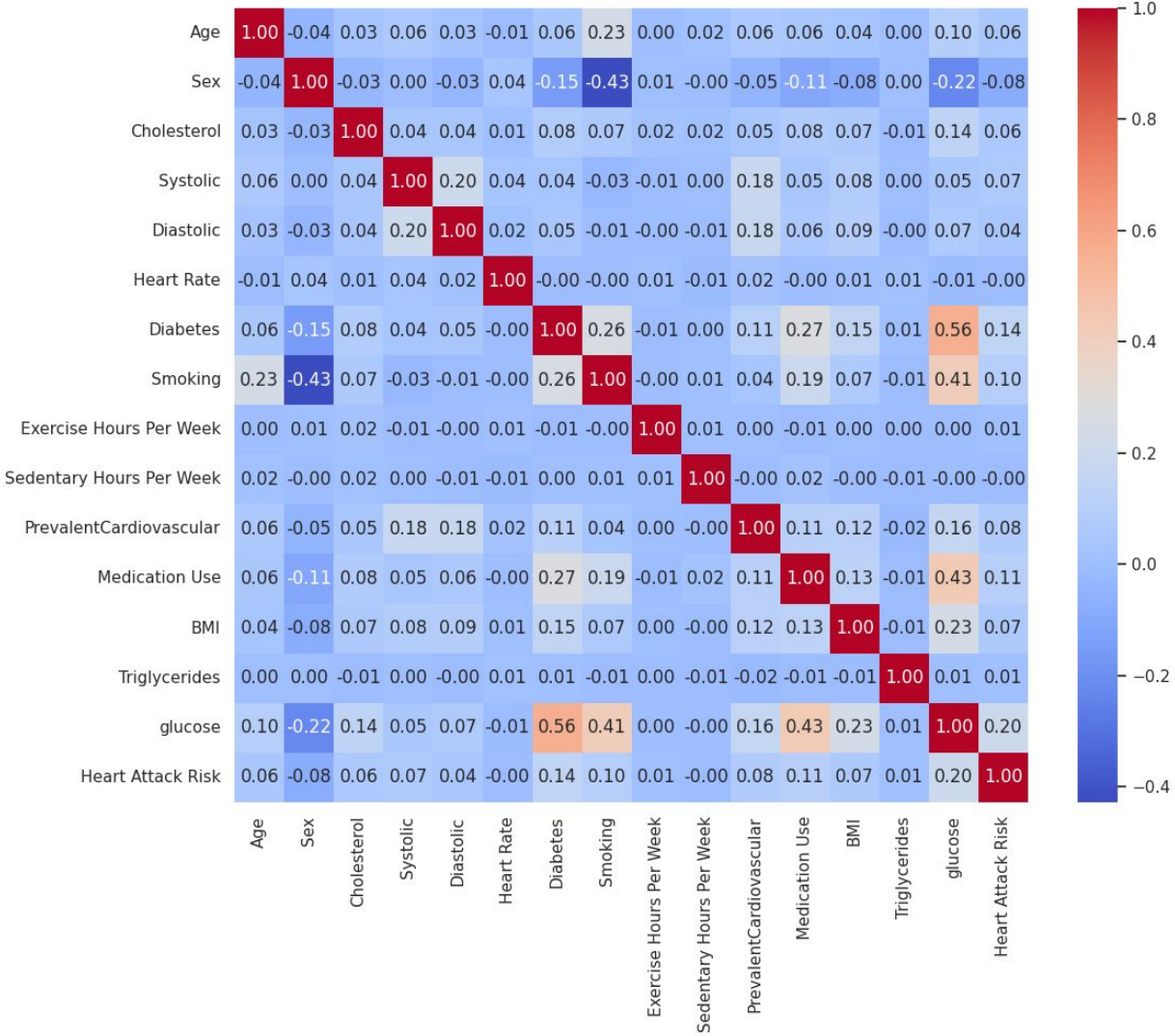
Análisis bivariado : Gluscosa y Diabetes

Insight : Los niveles de glucosa están relacionadas directamente a la condición de diabetes.



Como se imputaron los faltantes?

A través de la media según la zona de diabetes en la que se encontrara el paciente.



Correlación

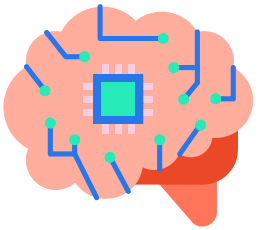
Alta correlación positiva:

- Glucose - Smoking
- Diabetes - Glucosa
- Medication use - Glucose
- Diabetes - Medication Use

Alta correlación negativa:

- Sex - Smoking
- Sex - Diabetes
- Diabetes - Smoking

Recomendaciones poblacionales



Factores Demográficos

- Enfocar las campañas en **hombres** y **personas mayores de 40 años** para reducir la prevalencia de factores de riesgo de enfermedades cardíacas.
- Promover **estilos de vida saludables** y el **control regular de salud** en estos grupos de mayor riesgo.

Colesterol y Trigl.

- Implementar **programas de intervención** para manejar los niveles de colesterol y triglicéridos elevados.
- **Recomendaciones Nutricionales** y **medicamentos** según las necesidades individuales para mantener los niveles dentro de los límites saludables.

Diabetes vs glucosa

- Fomentar el **monitoreo regular de la glucosa** para identificar y manejar la pre diabetes la diabetes.
- **Educación sobre la gestión de la diabetes**, incluyendo cambios en la dieta y la actividad física, para mantener los niveles de glucosa bajo control.

Etapas de Modelado

01

Modelos de clasificación con parámetros predeterminados

Resultados no motivadores.

02

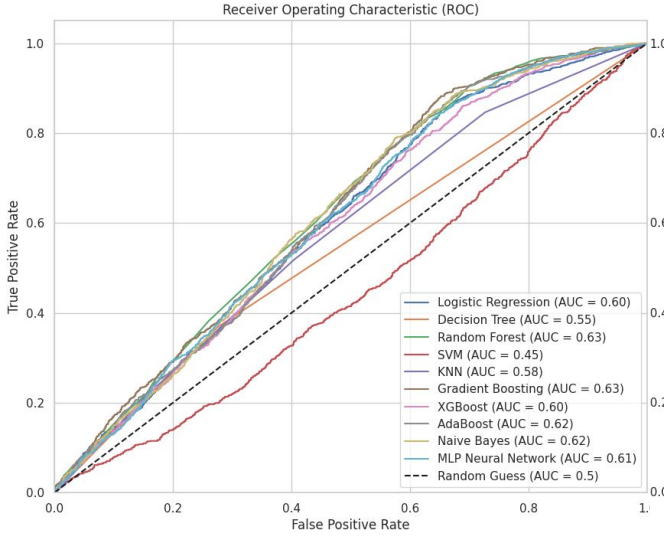
Balanceo de datos

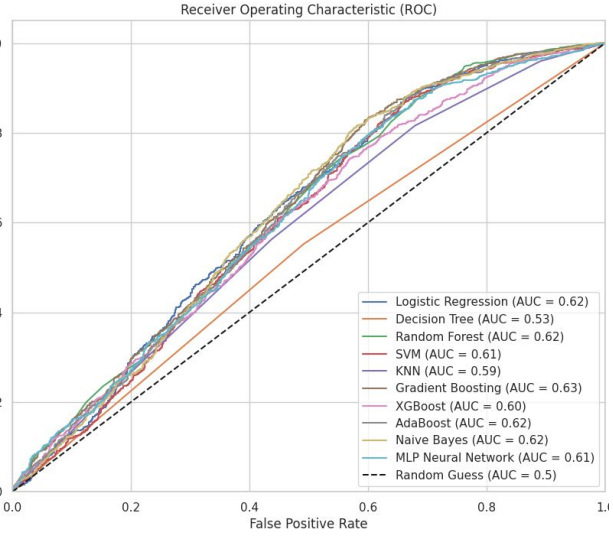
Se realiza un subsampling del grupo mayoritario para balancear las etiquetas.

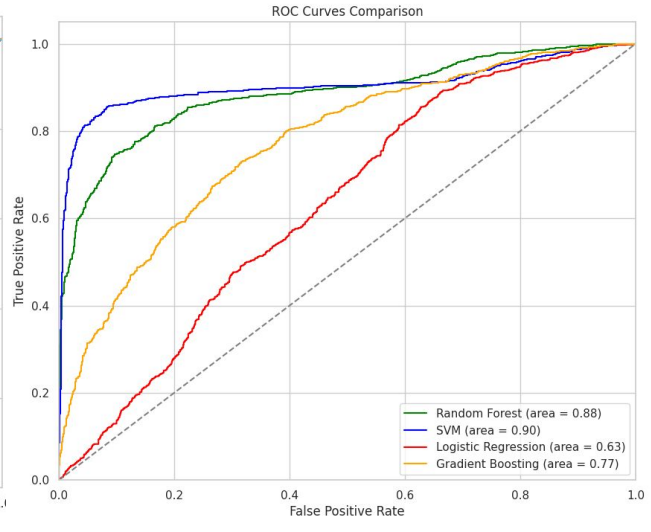
03

Selección de modelos , optimización y validación

A través de RandomizedSearchCV y GridSearchCV .







Que analizamos para comparar los modelos?

Del classification Report nos interesa...

01

RECALL

Mide cuán bien está el modelo identificando a las personas con riesgo de ataque cardíaco. Con datos desbalanceados esta métrica te ayuda a asegurarte de que el modelo no está fallando en su predicción.
Es crucial donde es más importante detectar todos los casos positivos que minimizar los falsos positivos.

02

F1 - SCORE

Es la media armónica entre la precisión y el recall. Es útil cuando tienes un desbalance entre clases y quieres un **equilibrio entre la capacidad del modelo para detectar positivos y minimizar los falsos positivos.**

03

ACCURACY

Mide el porcentaje de predicciones correctas (positivas y negativas). Es intuitiva, pero **puede ser engañosa si las clases están desbalanceadas.**

04

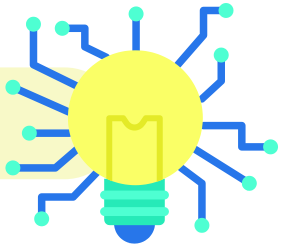
CURVA ROC - AUC

Mide la capacidad del modelo para distinguir entre las clases. **Un valor más alto indica mejor discriminación**.

Comparación de entrenamientos

Modelos entrenados	Recall		F1 - Score		Accuracy	ROC AUC
	0 (No riesgo de ataque cardiaco)	1 (Si riesgo de ataque cardiaco)	0 (No riesgo de ataque cardiaco)	0 (No riesgo de ataque cardiaco)		
Random Forest	0.49	0.90	0.61	0.74	0.69	0.88
Gradient Boosting	0.50	0.86	0.61	0.72	0.68	0.77
Logistic Regression	0.46	0.73	0.54	0.64	0.59	0.63
Support-vector machines	0.22	0.95	0.35	0.69	0.58	0.90

Cuales son los modelos que elegimos de nuestro entrenamiento ?



RANDOM FOREST

Ventajas:

- ➔ Buen equilibrio entre precision y recall, especialmente en la clase 1 (ataques cardíacos).
- ➔ La alta AUC sugiere que el modelo es bastante efectivo en distinguir entre las clases.

Desventajas:

- ➔ Aunque la precisión y recall en la clase 0 son moderadas.

SVM

Ventajas:

- ➔ Alta AUC
- ➔ Buen porcentaje de Recall

Desventajas

- ➔ Muestra un recall bajo en la clase 0, no identifica bien los casos negativos. Esto podría ser preocupante dependiendo del costo de los falsos positivos.

Algunos problemas durante el proceso..

01 **Distribución de Variables**

Fue necesario complementar el dataset original para aplicar variabilidad a los datos, e imputar los faltantes evaluando cuidadosamente cada caso.

02 **PCA**

Utilizarlo concluye en una pérdida significativa de la información.

03 **Desbalanceo de datos**

Se resolvió entrenando con un subsampling de la clase mayoritaria.

04 **Validación cruzada y K-Fold**

No fue óptimo para todos los modelos.

05 **Escalado de datos.**

No fue necesario en todos los modelos.