Heart Disease Prediction

PREDICCIÓN DE ENFERMEDADES CARDÍACAS

VALERIA MICOL GARCÍA

Heart Disease

CUIDEMOS NUESTRA SALUD CARDÍACA

CONTEXTO

Según la Organización Mundial de la Salud, las enfermedades cardiovasculares son la primer causa de muerte del mundo, siendo la responsable de la pérdida de 18 millones de vidas al año

OBJETIVO

Se tiene la intención de entrenar un algoritmo para detectar cuáles son los factores que influyen en la posibilidad de contraer una enfermedad cardíaca

AUDIENCIA

Este modelo puede ser utilizado por personal de la salud como también por personas interesadas en su bienestar general

Preparación de datos

1. Hipótesis y Objetivo
2. Data Adquisition
3. Data Wrangling
4. Análisis Univariado
5. Análisis Bivariado
6. Pasos previos al entrenamiento

Entrenamiento

7 Balanceo variable target

8 Análisis PCA

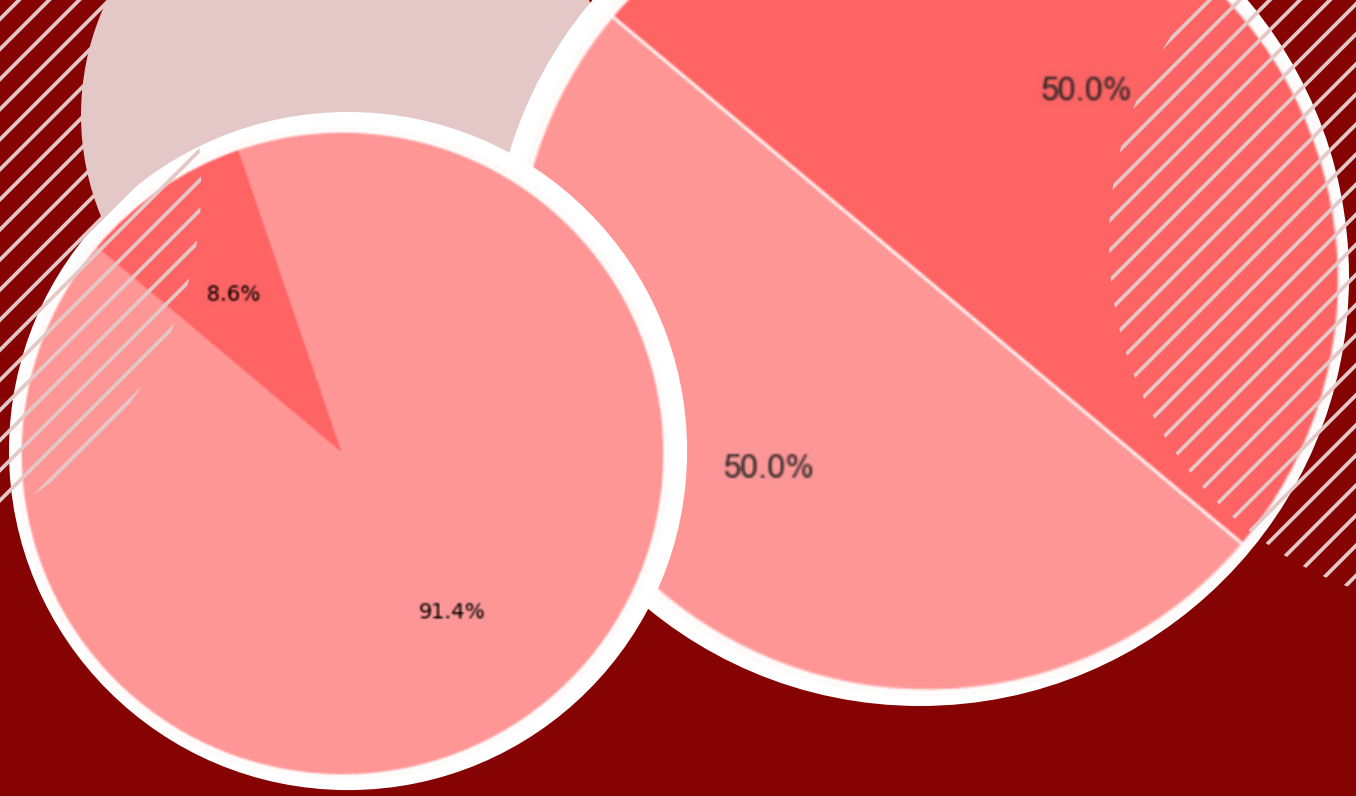
9 Árbol de decisión

10 Regresión Logística

11 Random Forest

12 XGBoosting

12 Comparación de modelos



Balanceo de la variable target

PROBLEMA

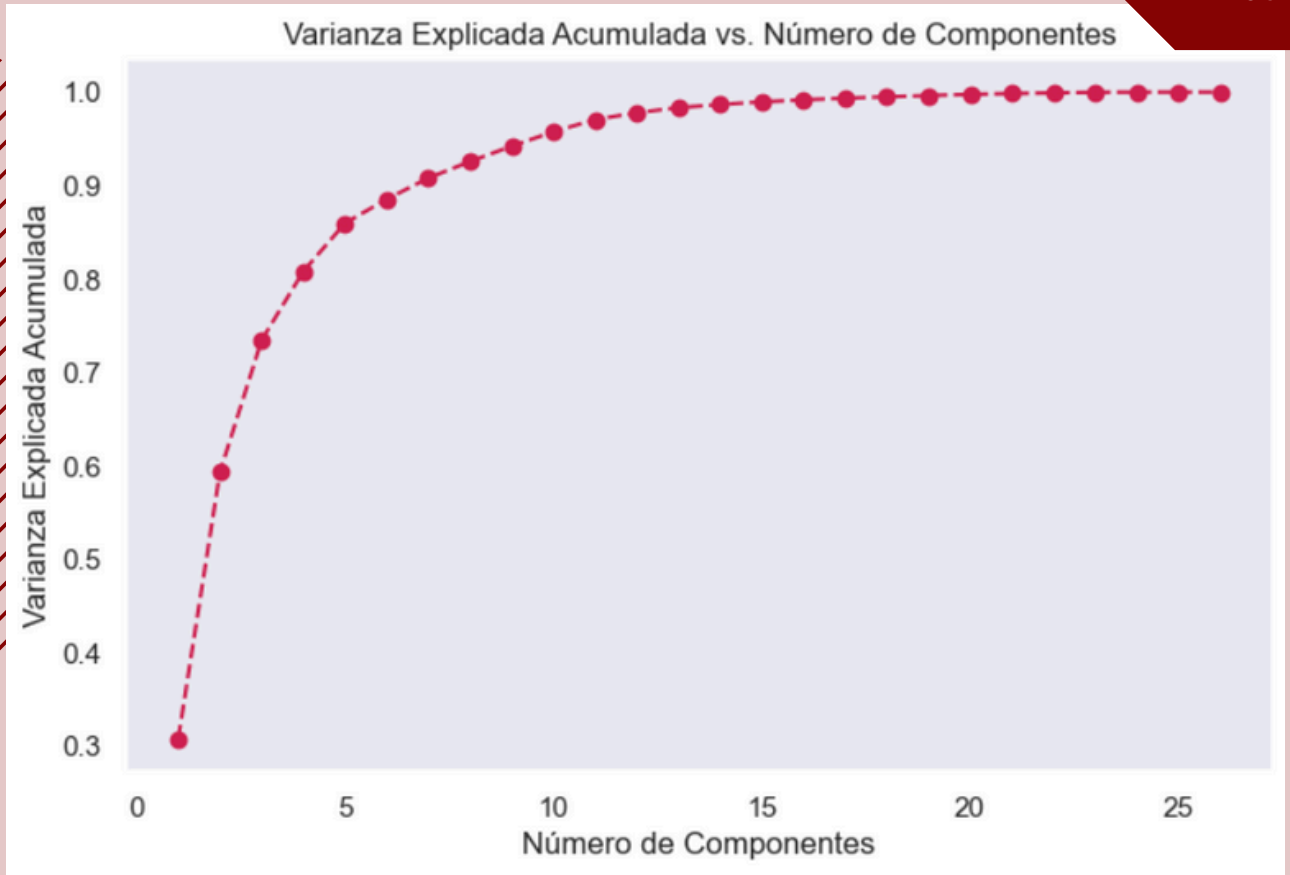
Se observó que la variable objetivo estaba desbalanceada, con algunas clases significativamente menos representadas. Este desbalance puede afectar negativamente el rendimiento del modelo, sesgándolo hacia la clase mayoritaria y dificultando el aprendizaje de la clase minoritaria.

SOLUCIÓN

Para resolver este problema, se utilizó una técnica de sobremuestreo, el RandomOverSampler, que genera nuevas instancias para la clase minoritaria y equilibra las clases en el conjunto de entrenamiento. Esto ayuda al modelo a aprender mejor y a predecir con mayor precisión.

Análisis PCA

Menor número de dimensiones implica menor tiempo de entrenamiento y menor recurso computacional y mayor performance



método del codo:

Con el método del codo, puedo ver que a partir del componente 15, la varianza se incrementa de forma mínima, por lo que tomo ese número para aplicar pca

Se llevó a cabo un análisis de Componentes Principales (PCA) y se aplicó a todos los modelos evaluados. Sin embargo, los resultados no mostraron mejoras en las métricas de rendimiento. Por ello, en el notebook final se optó por no utilizar las componentes principales en la mayoría de los modelos. La excepción fue el árbol de decisión, para el cual las métricas de rendimiento permanecieron equivalentes con y sin la aplicación de PCA.

ÁRBOL DE DECISIÓN

ACCURACY | 0.69

PRECISION | 0.91

RECALL | 0.69

F1-SCORE | 0.76

REGRESIÓN LOGÍSTICA

ACCURACY | 0.74

PRECISION | 0.90

RECALL | 0.73

F1-SCORE | 0.79

RANDOM FOREST

ACCURACY | 0.73

PRECISION | 0.90

RECALL | 0.73

F1-SCORE | 0.78

XGBOOSTING

ACCURACY | 0.69

PRECISION | 0.91

RECALL | 0.69

F1-SCORE | 0.76

Conclusión

	Métrica	Decision Tree	Random Forest	Regresión Logística	XGBoost
0	Accuracy	0.692666	0.732000	0.740656	0.693429
1	Precision	0.910052	0.907695	0.907183	0.910128
2	Recall	0.692666	0.732000	0.740656	0.693429
3	F1 Score	0.760404	0.789734	0.796008	0.760989

Comparando todos los modelos, se observa que la precisión es bastante alta, alcanzando un 90%, con el DecisionTree y el XGBoosting llegando incluso al 91%. Sin embargo, en términos de recall, accuracy y F1 Score, el RandomForest y el modelo de Regresión Logística muestran métricas más sólidas. Aunque las diferencias son mínimas, la elección final se inclina hacia el modelo de Regresión Logística, que supera ligeramente al RandomForest en rendimiento general.

MODELO FINAL → **REGRESIÓN LOGÍSTICA**

GITHUB

En el siguiente link se accede al proyecto:

https://github.com/valemicolgarcia/Heart-Disease-Prediction

ALUMNA: VALERIA MICOL GARCÍA

TUTOR: MATEO BONGIORNO

PROFESOR: GERMÁN RODRIGUEZ