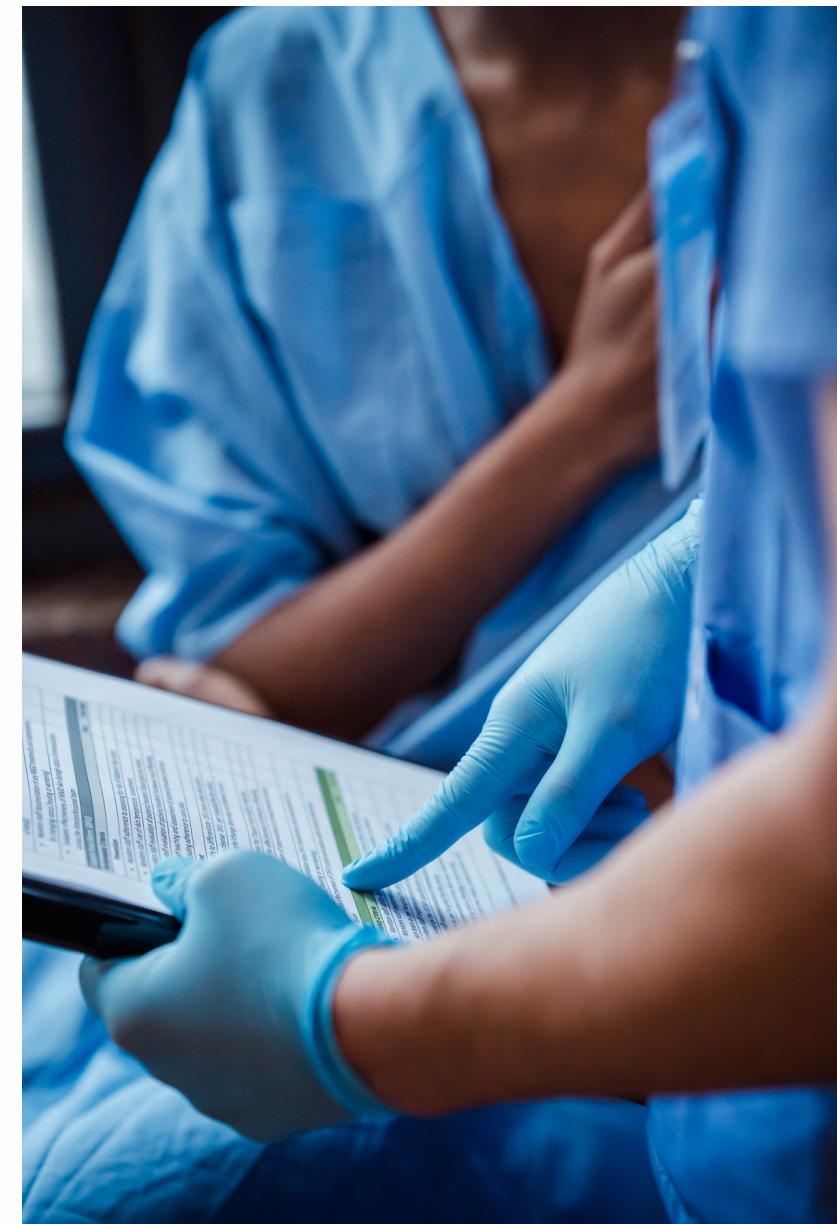


PREDICCIÓN DEL RIESGO CARDIOVASCULAR



ÍNDICE

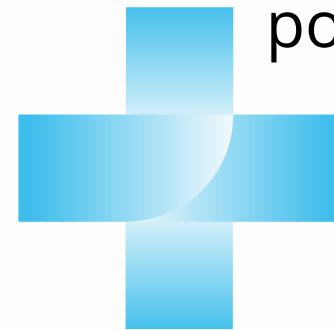
- 1 Introducción
- 2 Objetivo
- 3 EDA
- 4 Modelado
- 5 Selección del modelo
- 6 Conclusiones



INTRODUCCIÓN

Hoy en día las encuestas de salud son una parte fundamental del sistema de salud de cualquier país, y ayudan a la toma de decisiones relativas a las políticas de salud.

Los modelos supervisados de Machine Learning permiten, a partir de estas encuestas, detectar patrones y predecir posibles patologías.



OBJETIVO

Desarrollar un modelo de Machine Learning que prediga el riesgo de enfermedad cardíaca en la población, a partir de características del sujeto, hábitos de vida y estado de salud, y la concurrencia con otras patologías.

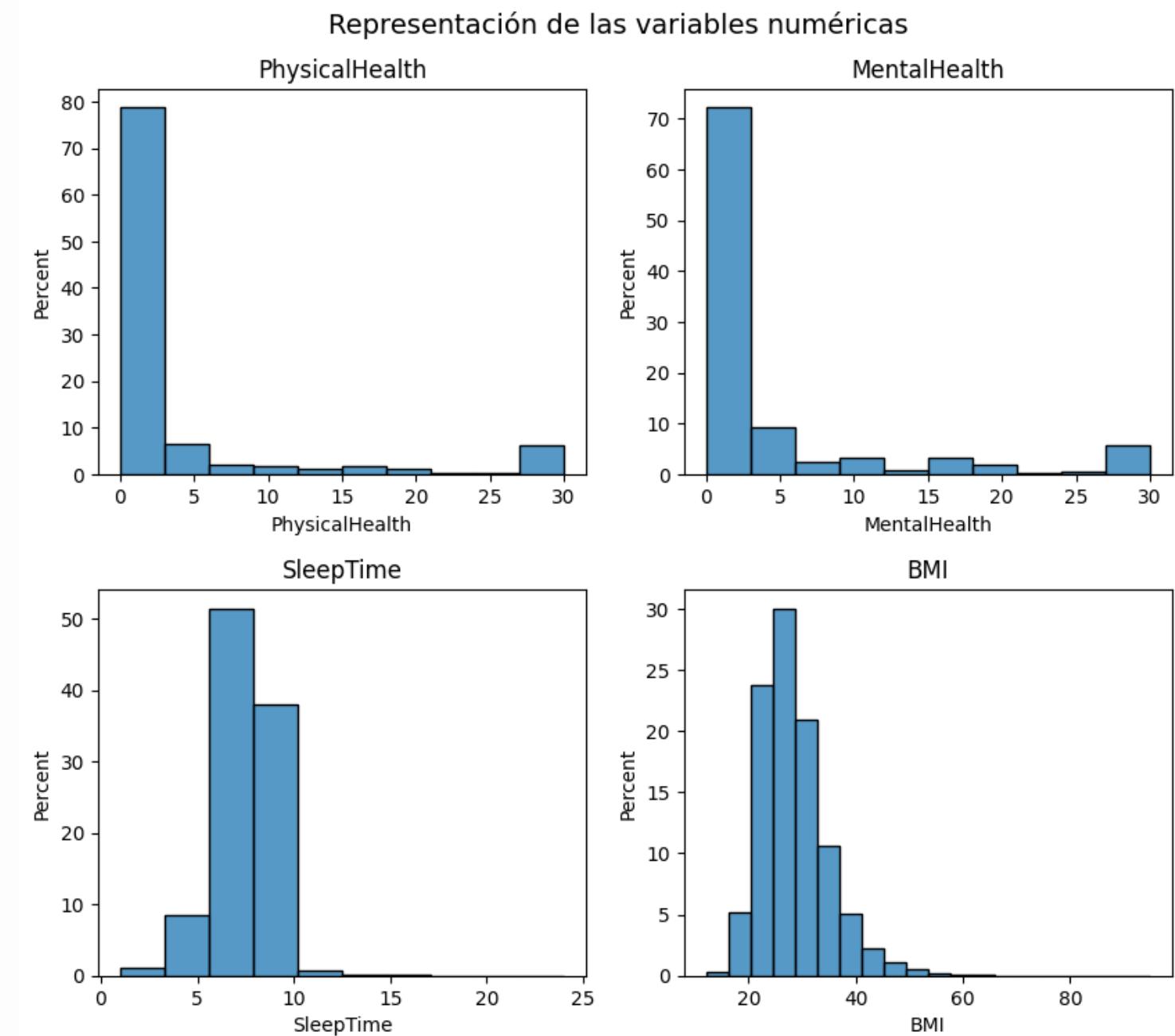
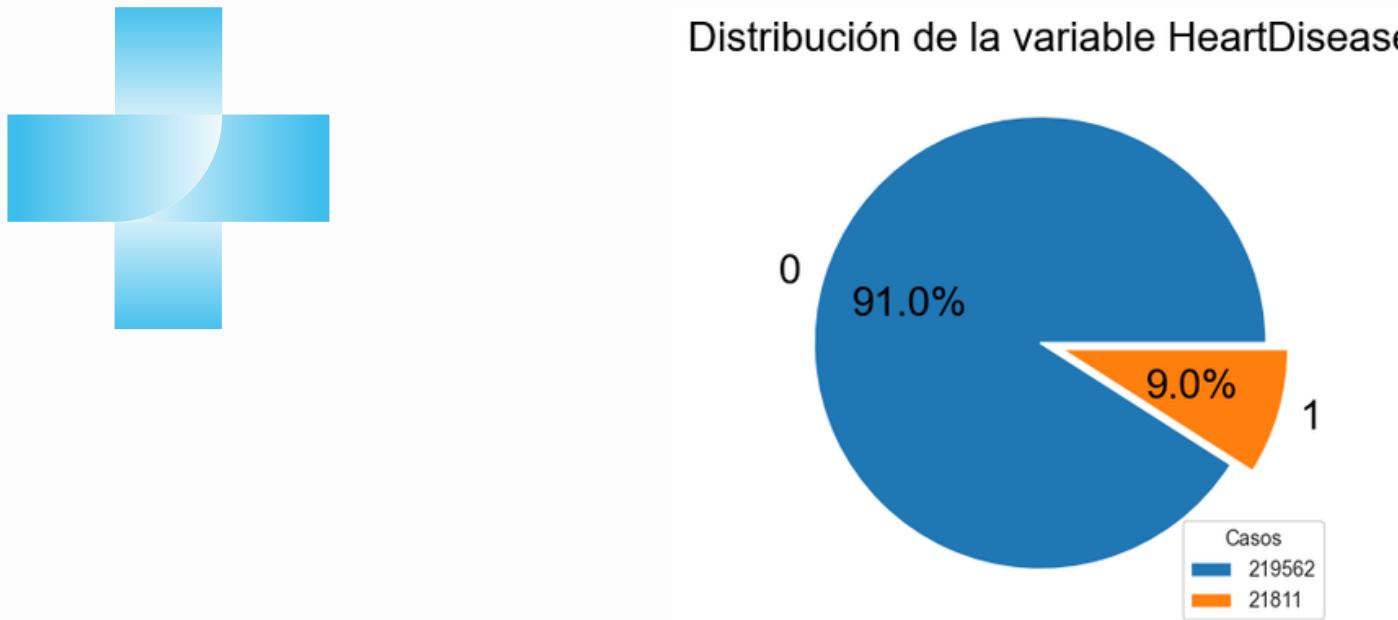
Los datos han sido extraídos de la encuesta de salud anual realizada por Behavioral Risk Factor Surveillance System (BRFSS).



EDA

DESCRIPCIÓN DE LOS DATOS

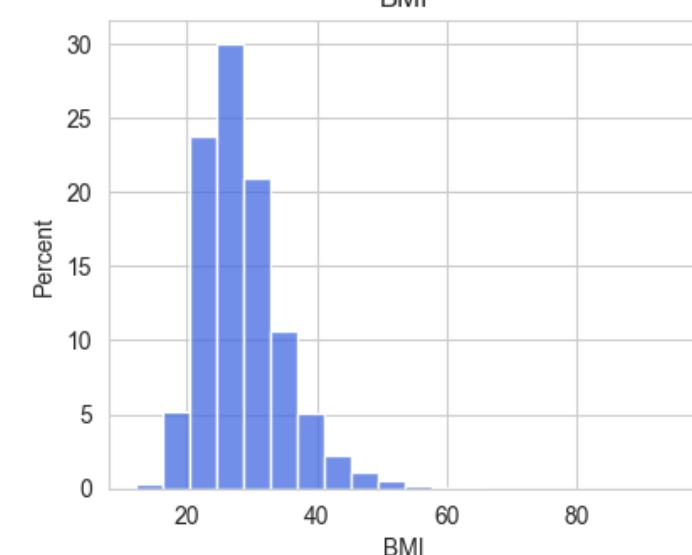
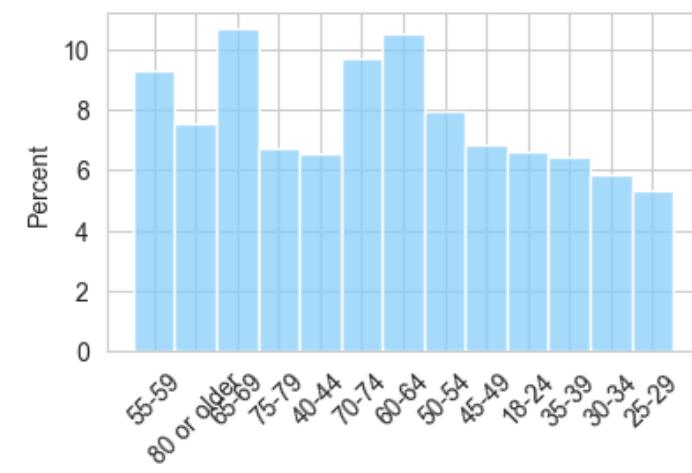
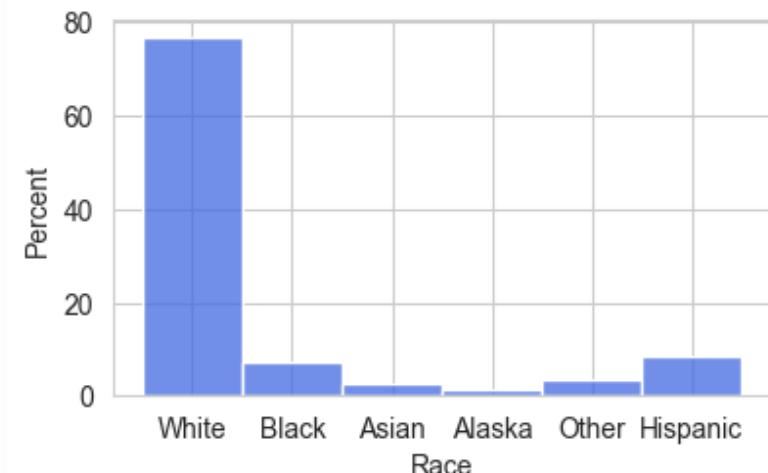
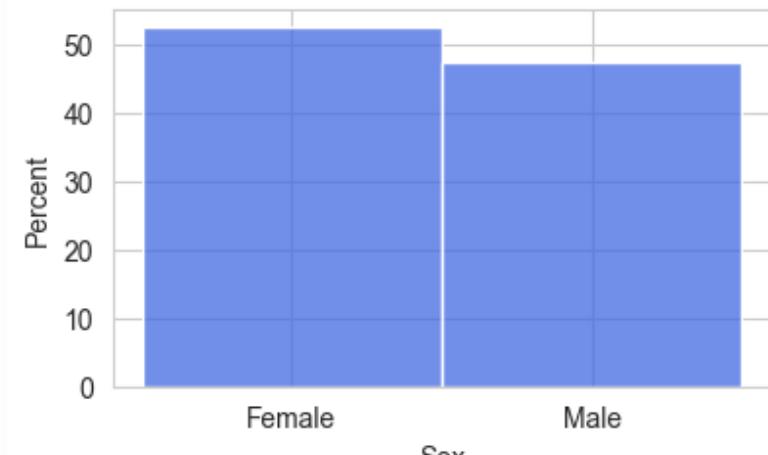
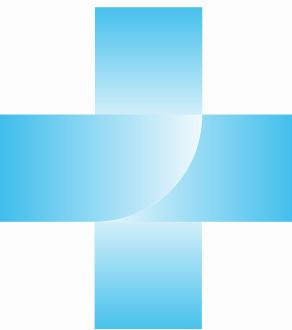
- 320.000 datos y 18 variables
- Conjunto de datos no balanceado: Variable target, 9% del total
- Ausencia de distribuciones normales



EDA

VARIABLES

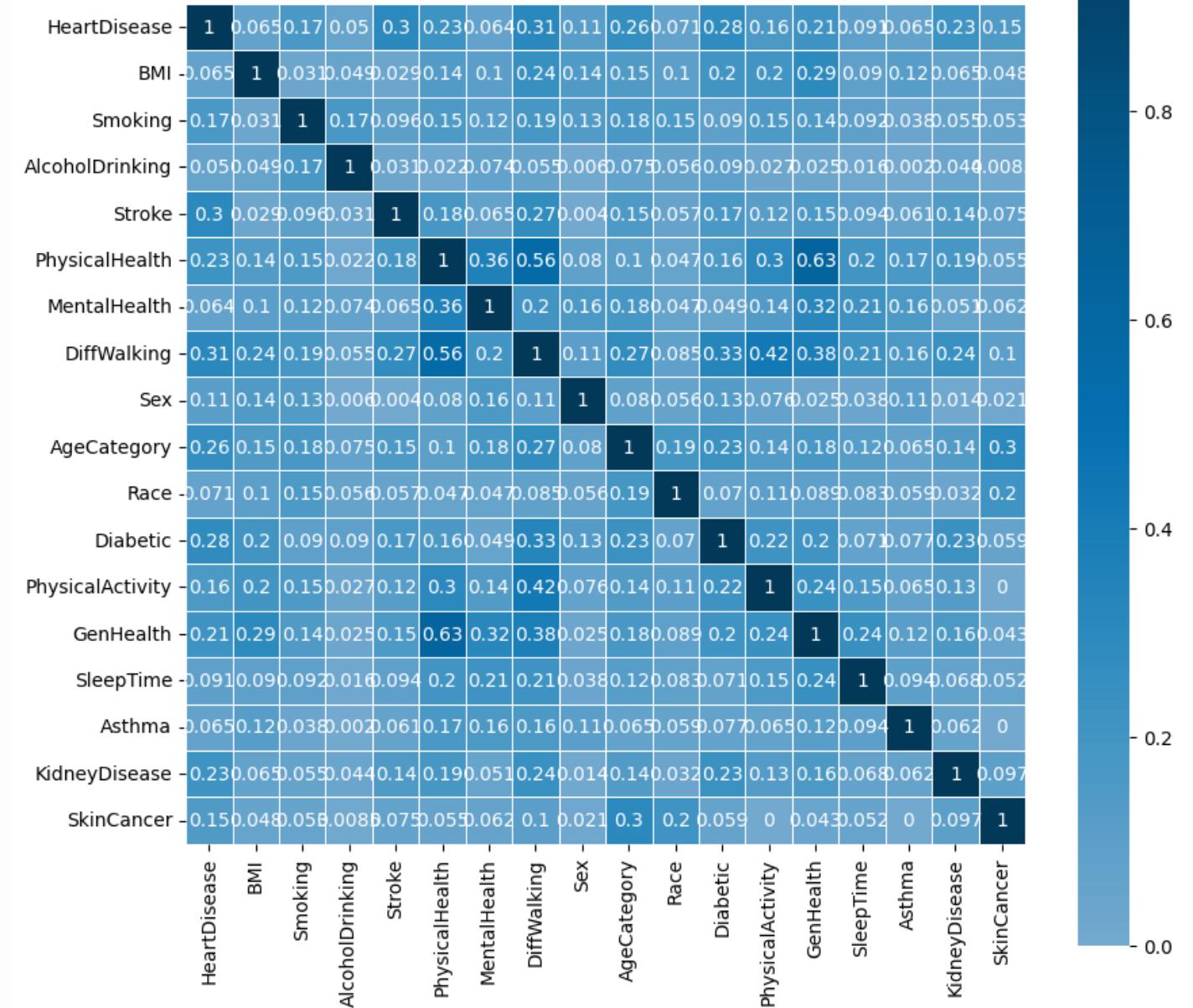
- Edad
- Género
- Índice de masa corporal
- Tabaquismo
- Estado general de salud
- Hábitos: ejercicio físico, horas de sueño, ingesta de alcohol
- Otras enfermedades: diabetes, afecciones renales, cancer de piel, asma
- Raza
- **Target: Enfermedad cardíaca**



EDA

CORRELACION

- Escasa correlación con la variable target
 - DiffWalking: 0.31
 - Stroke: 0.3
 - Diabetic: 0.28
 - AgeCategory: 0.26
 - PhysicalHealth: 0.23
 - KidneyDisease: 0.23

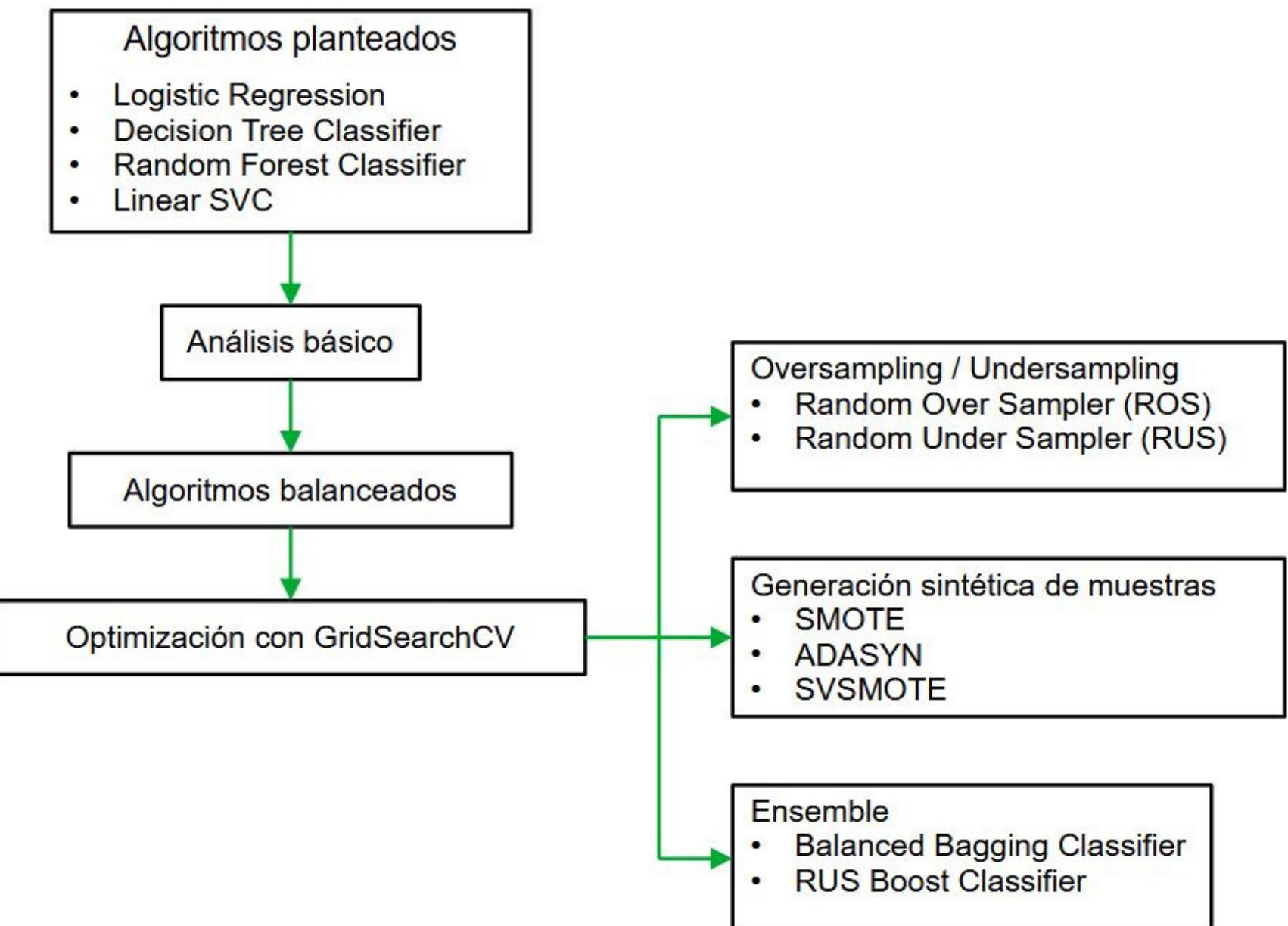
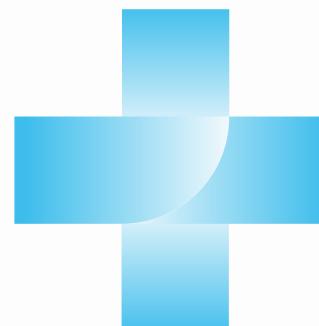


MODELADO

ESTRATEGIA

Estrategia distinta para datos no balanceados

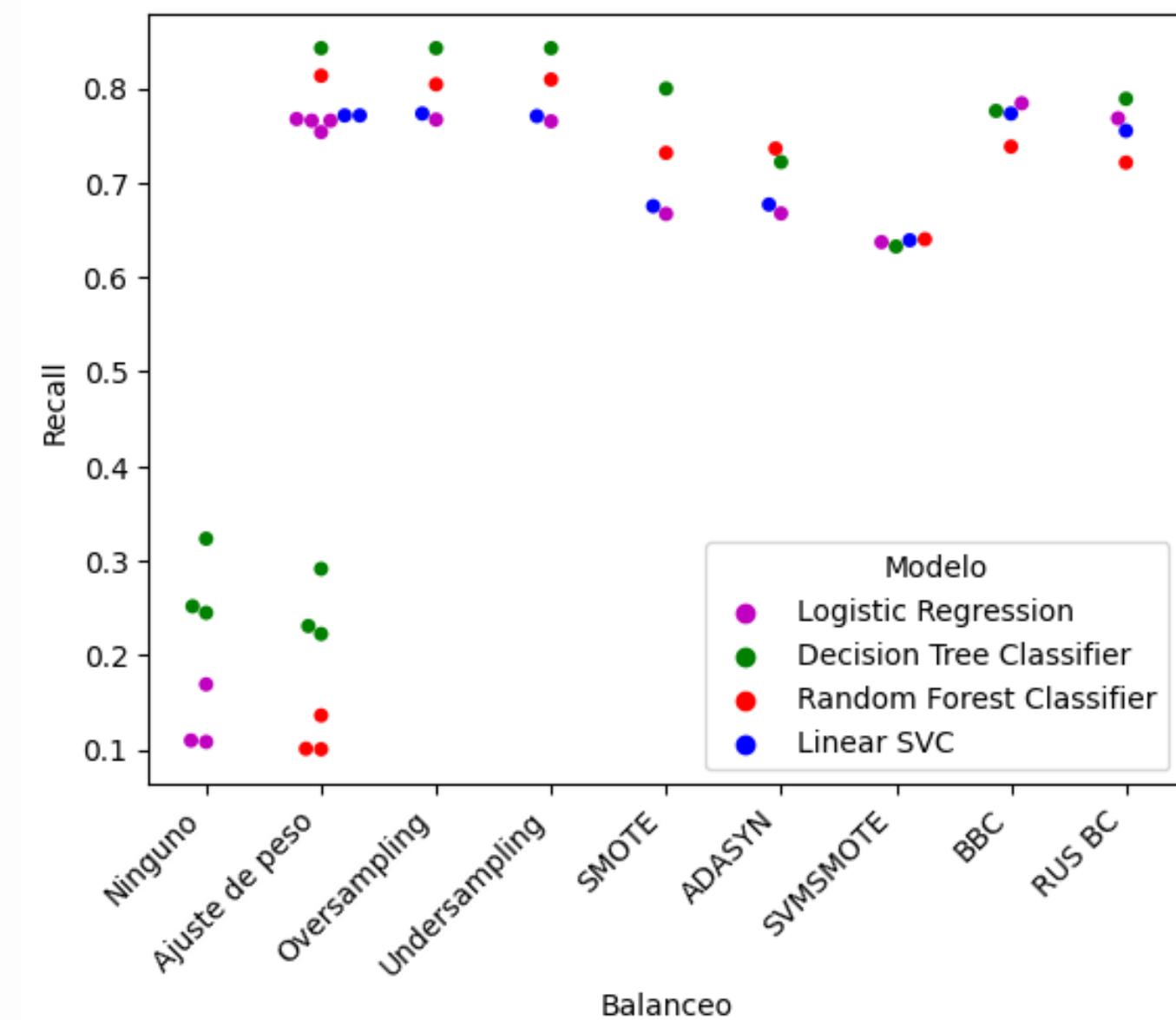
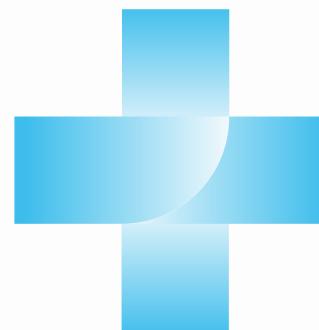
- Ajuste de peso
- Oversampling/Undersampling
- Generación sintética de muestras
- Ensemble



MODELADO

RESULTADOS

- Es necesario aplicar alguna estrategia de balanceo de los datos
- Undersampling y oversampling dan el mismo recall que el ajuste de peso
- Generación sintética de muestras da valores inferiores, al igual que los ensembles con Balanced Bagging Classifier y RUS Boosting Classifier

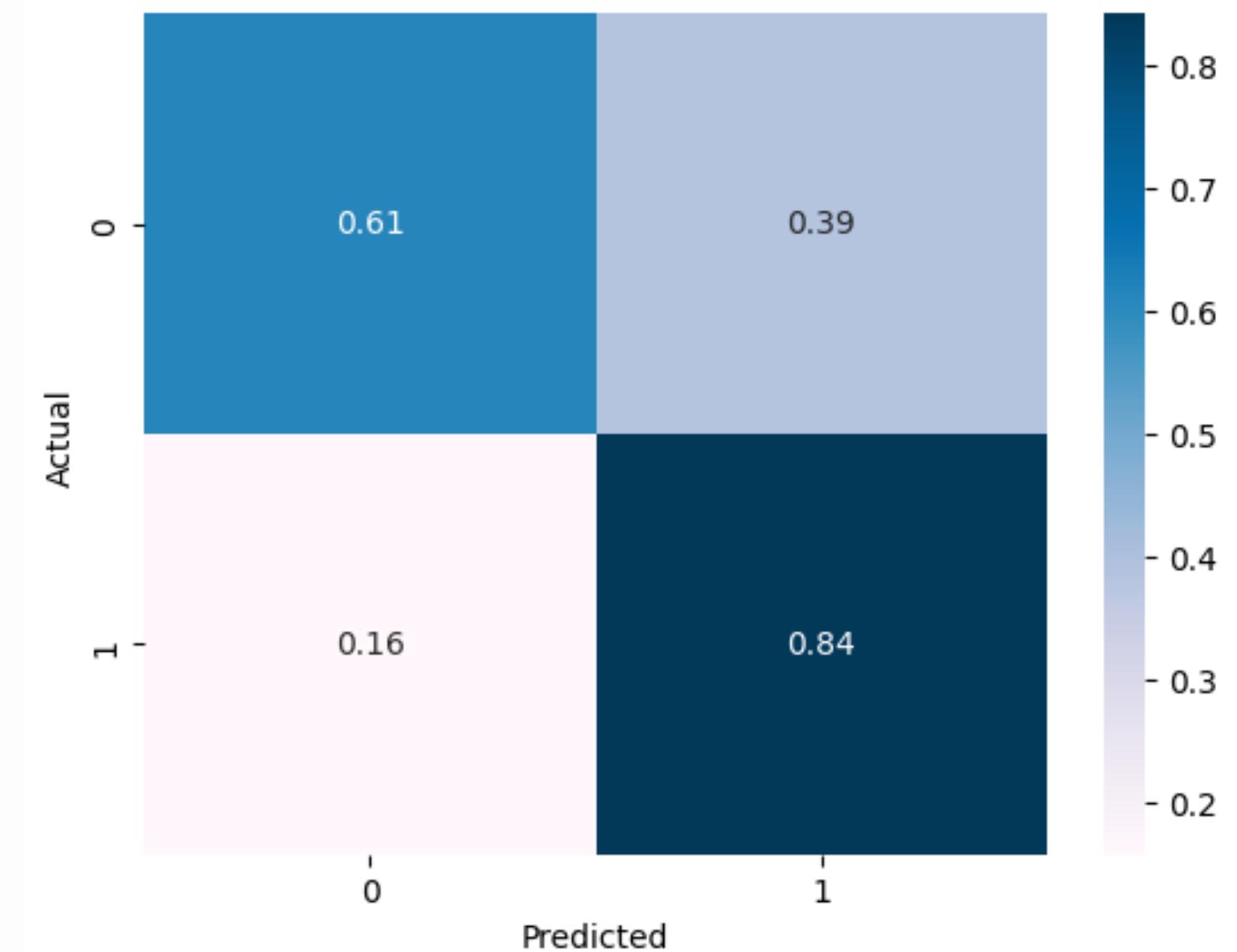
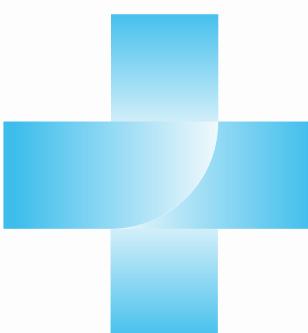


MODELADO

SELECCIÓN

DECISION TREE CLASSIFIER

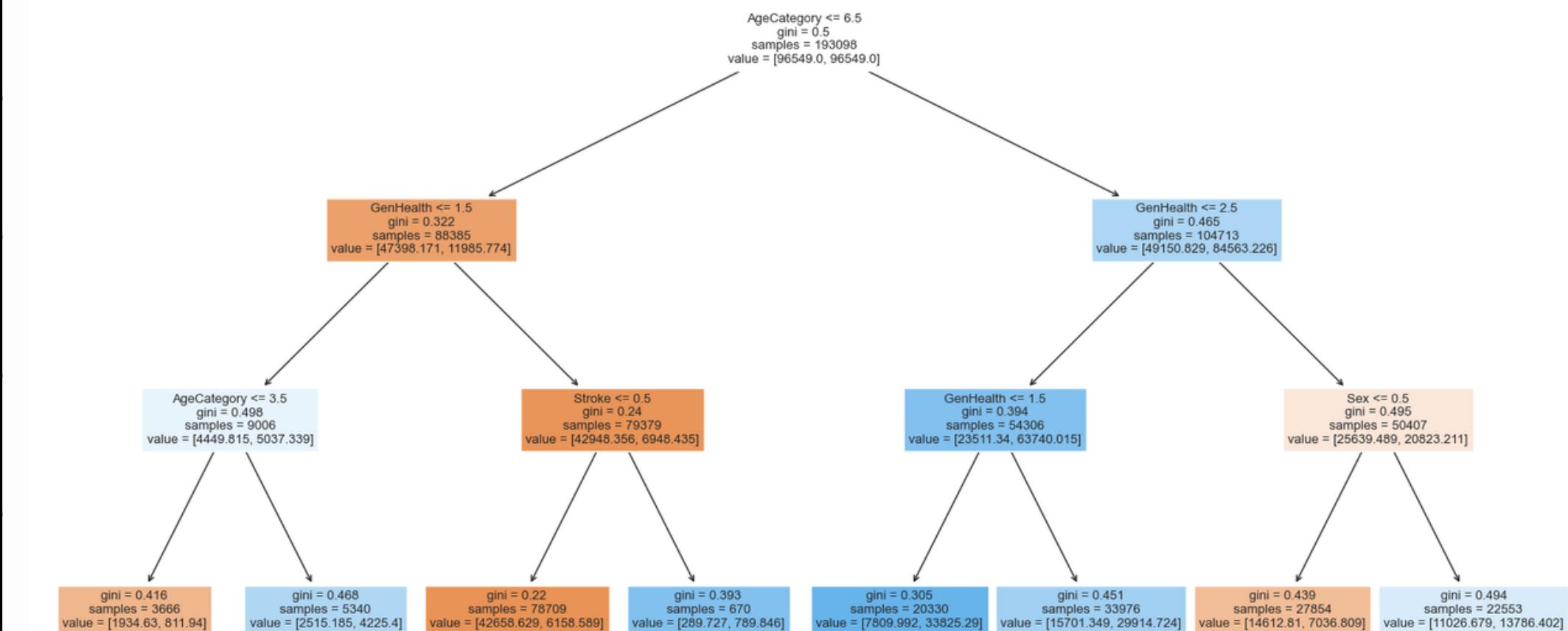
Recall	0.8422
class_weight	balanced
max_depth	3
min_samples_split	2



SELECCION DEL MODELO

FEATURE IMPORTANCE

Variable	Valor
Edad	0.602
Estado de salud	0.320
Género	0.047
Accidente cerebrovascular	0.029





CONCLUSIONES

- Importancia de trabajar con datos balanceados
- Variables más significativas:
 - Edad
 - Estado de salud
 - Género
- Variables que a priori se consideraban importantes, no se encuentran en el modelo seleccionado:
 - Tabaquismo
 - Ingesta de alcohol
 - Índice de masa corporal
- Puntos de mejora
 - Reformulación de las preguntas
 - Añadir en el estudio la hipercolesterolemia y la hipertensión
 - Obtener un conjunto de datos más balanceado
 - Eliminar la variable *Stroke*



**GRACIAS POR
VUESTRA ATENCIÓN**