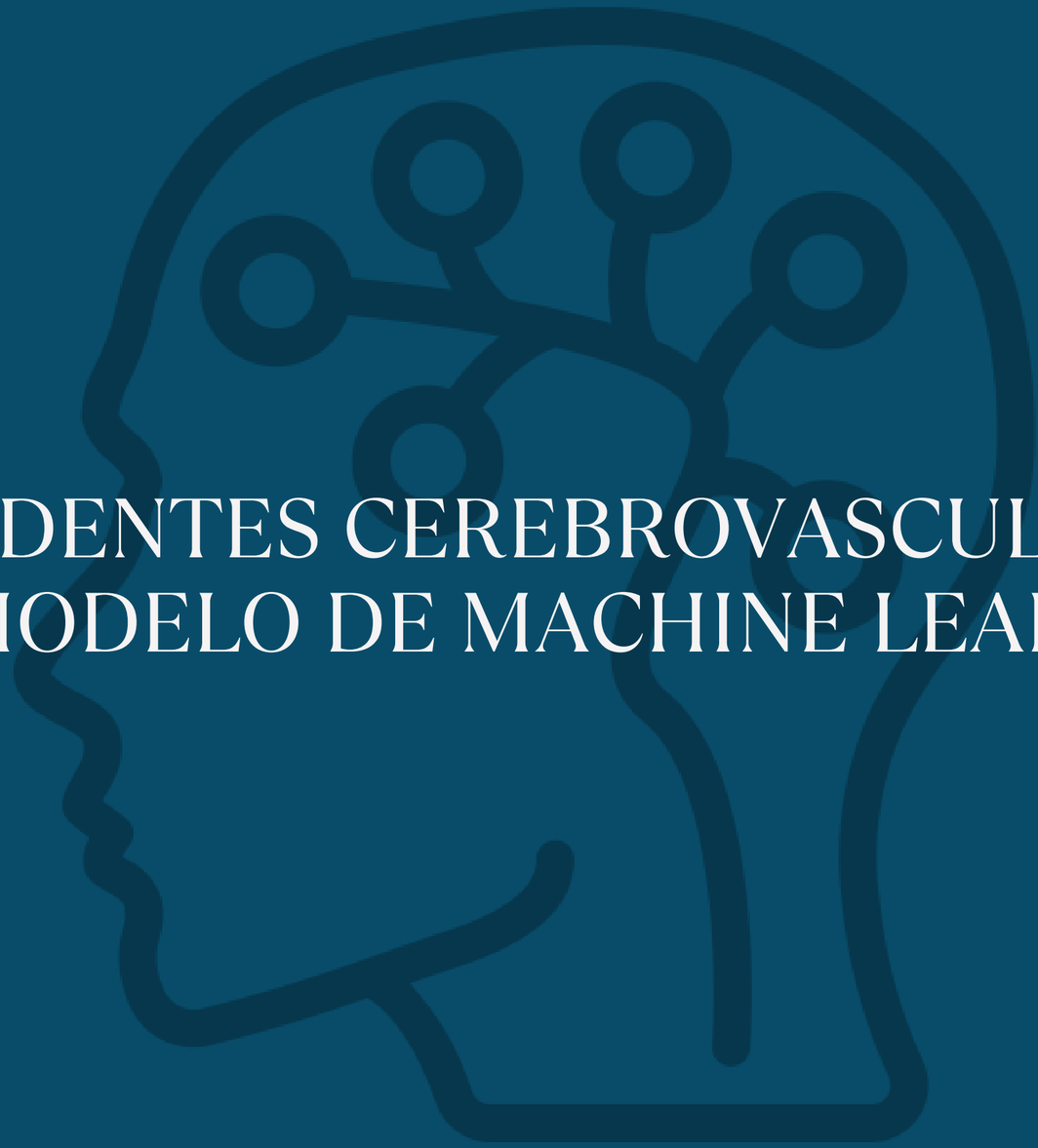
ACCIDENTES CEREBROVASCULARES
UN MODELO DE MACHINE LEARNING



CONTEXTUALICEMOS EL PROBLEMA

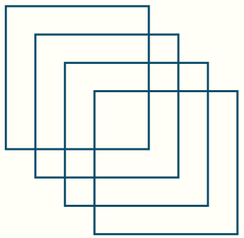
- Alta incidencia de casos
- Mortalidad elevada
- Altos costos asociados a su diagnóstico y tratamiento

NUESTROS OBJETIVOS CON MACHINE LEARNING

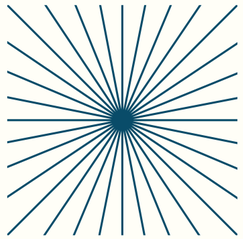
- Incitar a la consulta médica
- Mejor experiencia del paciente
- Prevención efectiva
- Optimización de recursos

FUENTE DE DATOS

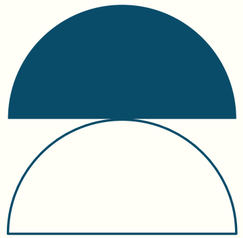
FUENTE DE DATOS



DETALLES



LIMPIEZA



FEATURES

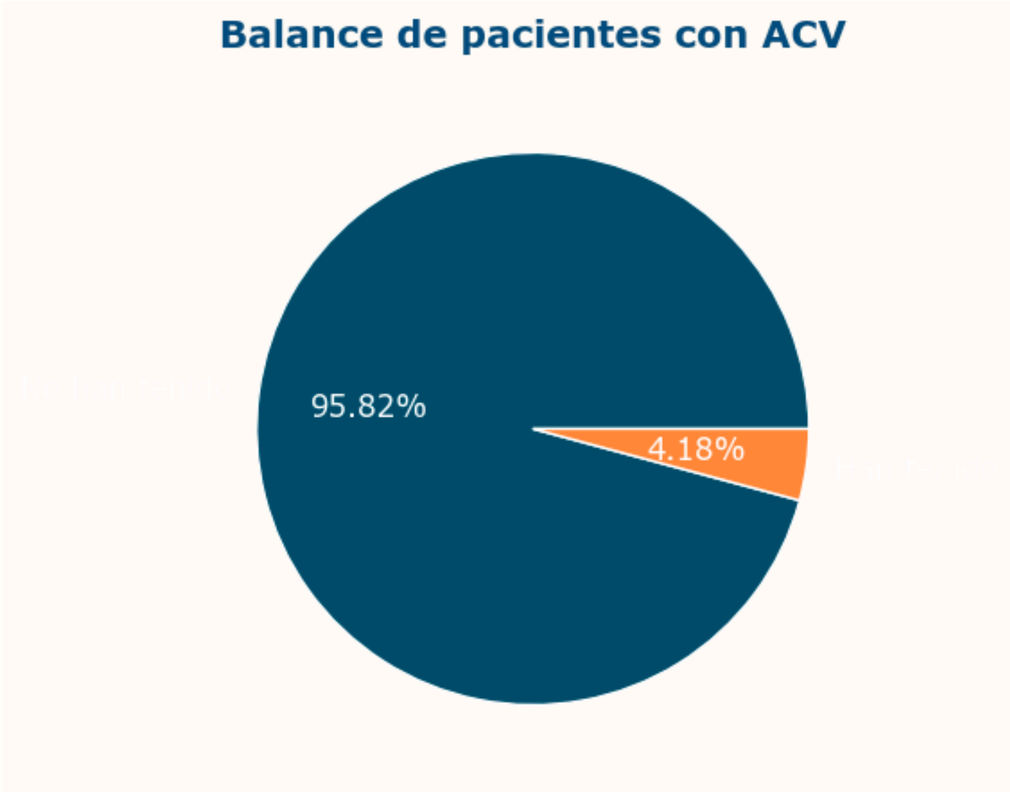
Información general

- Data of patients (Kaggle)
- 230.000 registros
- 34 columnas: edad, género, peso, enfermedades, etc.
- Sin valores nulos

DETALLES

Retos que nos presenta

- Selección de variables según nuestro contexto
- Desbalanceo de la variable a predecir



Eliminación de columnas

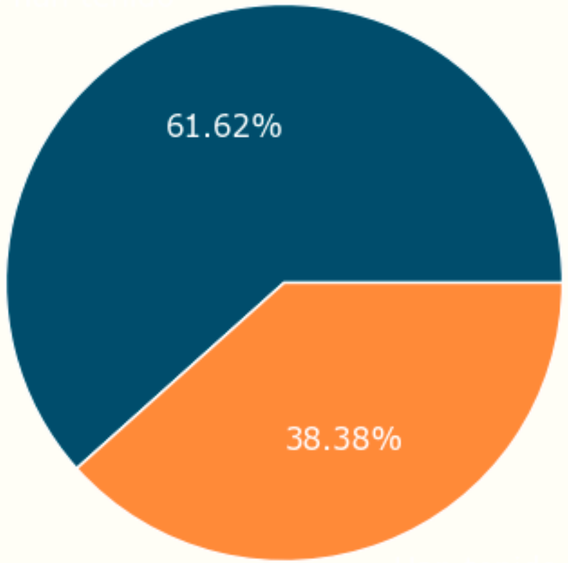
- COVID positivo
- HIV positivo
- Cancer de piel
- Ciudad
- etc.

Reducción del dataset

230000 registros -> 26000 registros

LIMPIEZA

Balance de pacientes con ACV



FEATURE ENGINEERING

Género

Mujer → 1

Hombre → 2

Rangos de edad

80 > ... → 4

60-79 → 3

40-59 → 2

18-39 → 1

Salud general

Excellent → 5

Very good → 4

Good → 3

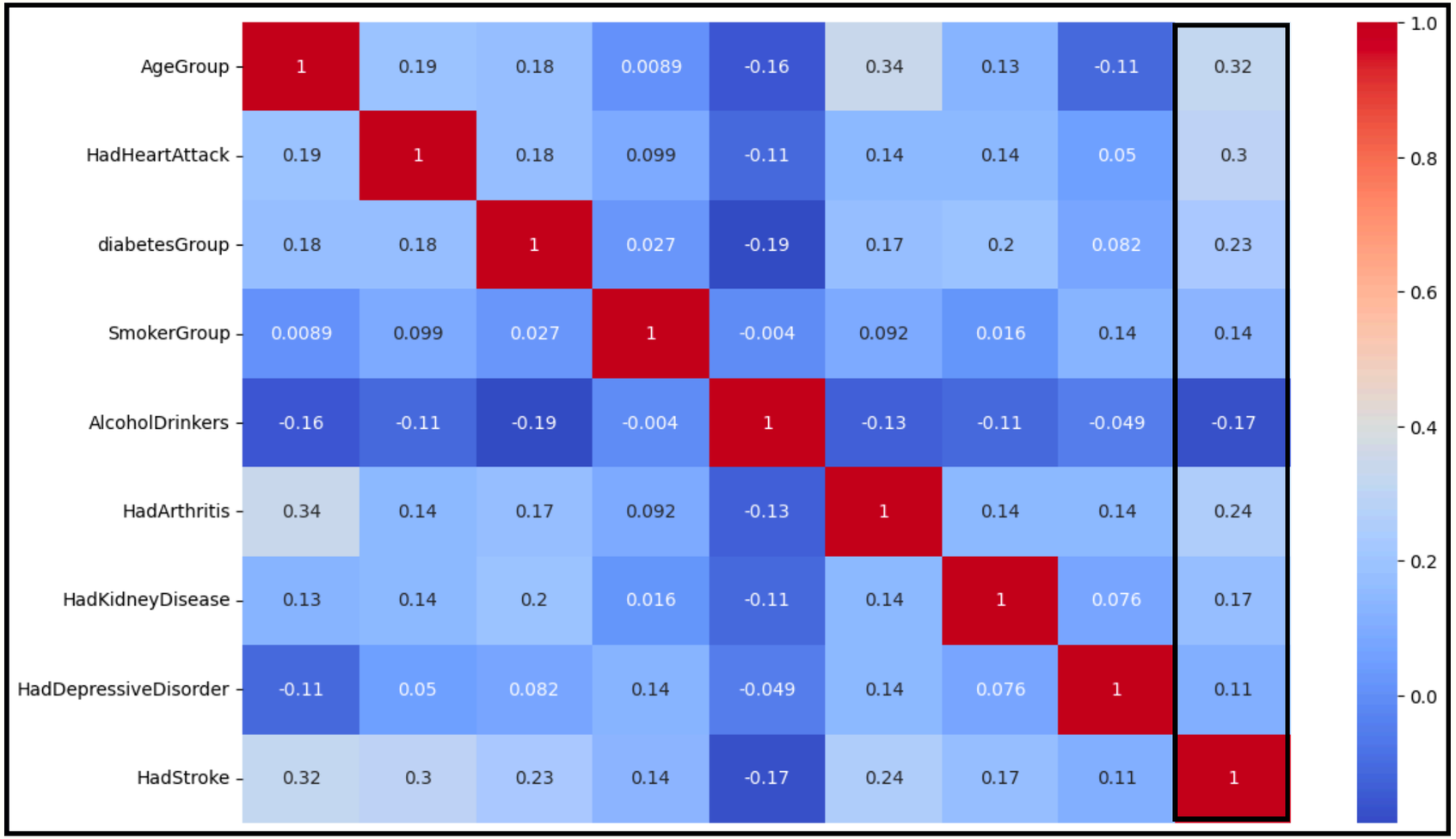
Fair → 2

Poor → 1

• • •

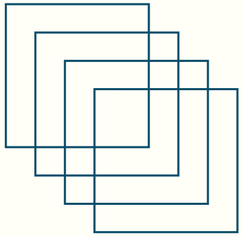
FUENTE DE DATOS

RELACIONES ENTRE VARIABLES



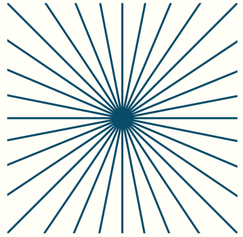
DESARROLLO DEL MODELO

RASGOS GENERALES



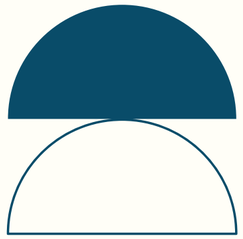
TRAIN-TEST

80%-20%



MÉTRICAS

F1-score
Precision
Recall



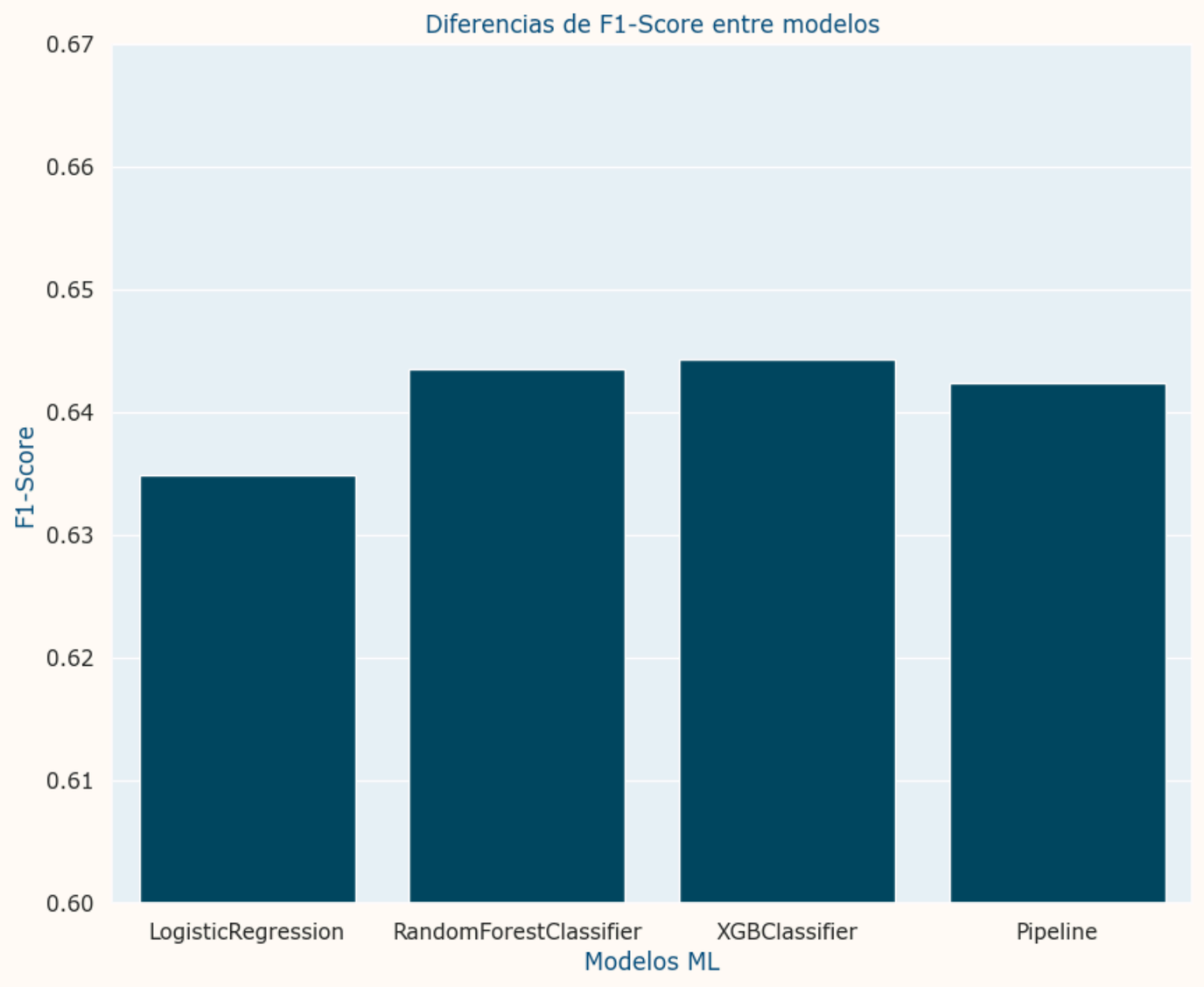
OBJETIVO

El modelo menos costoso en cuanto a tiempo y recursos.

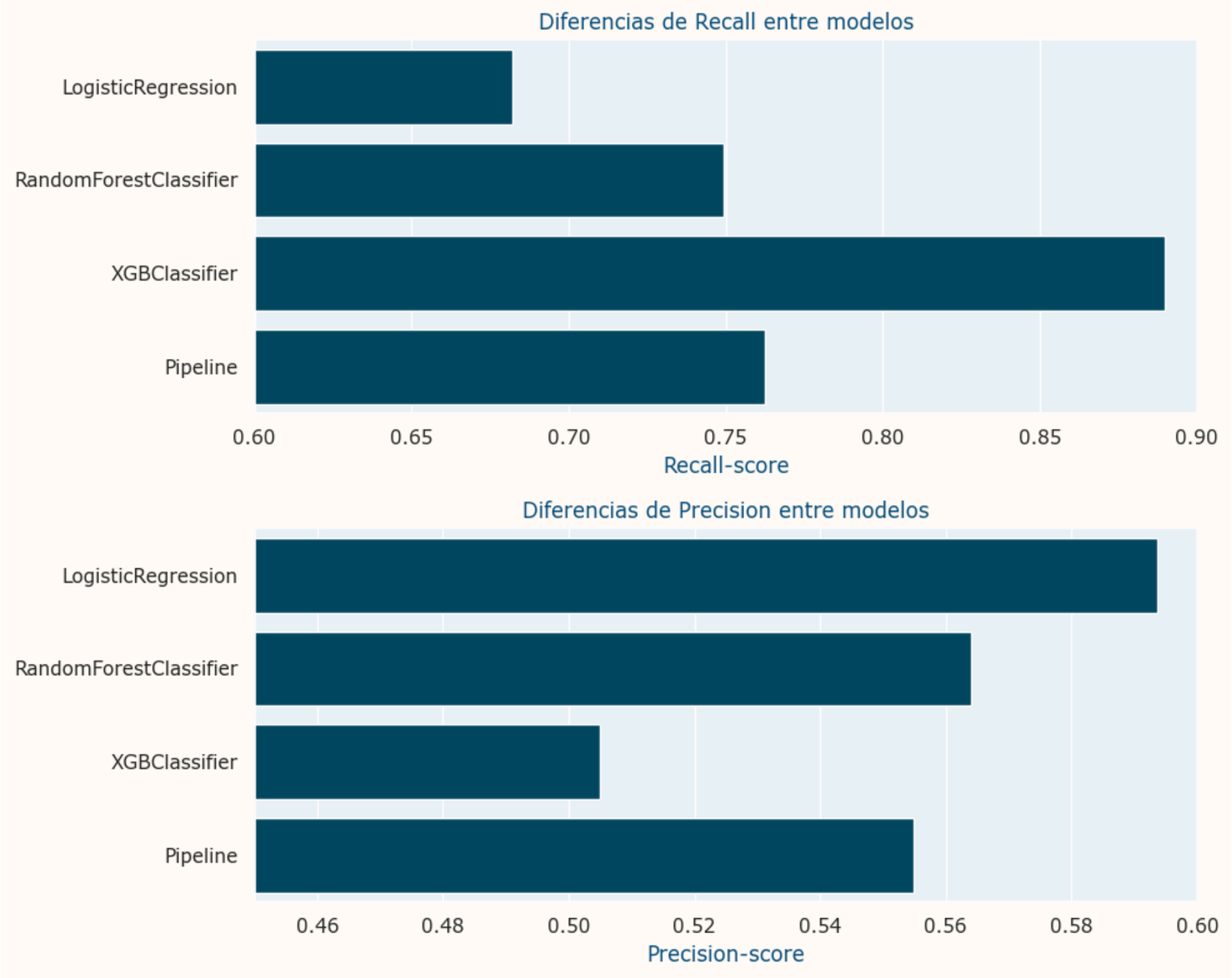
DESARROLLO DEL MODELO

LOGISTIC REGRESSION	RANDOM FOREST	XGBOOST	PCA-RANDOMFOREST
MAX_ITER = 1000 CLASS_WEIGHT = BALANCED	N_ESTIMATORS = 200 MIN_SAMPLES_SPLIT = 4 MAX_DEPTH = 6 CRITERION =GINI CLASS_WEIGHT = BALANCED	N_ESTIMATORS = 300 MIN_CHILD_WEIGHT = 3 MAX_DEPTH = 10 SCALE_POS_WEIGHT = 3 LEARNING_RATE = 0.05 GAMMA = 0.1	N_ESTIMATORS = 200 MIN_SAMPLES_SPLIT = 4 MAX_DEPTH = 6 CRITERION =GINI CLASS_WEIGHT = BALANCED N_COMPONENTS = 6
ACCURACY			
70.54%	68.84%	63.12%	68.12%

DESARROLLO DEL MODELO



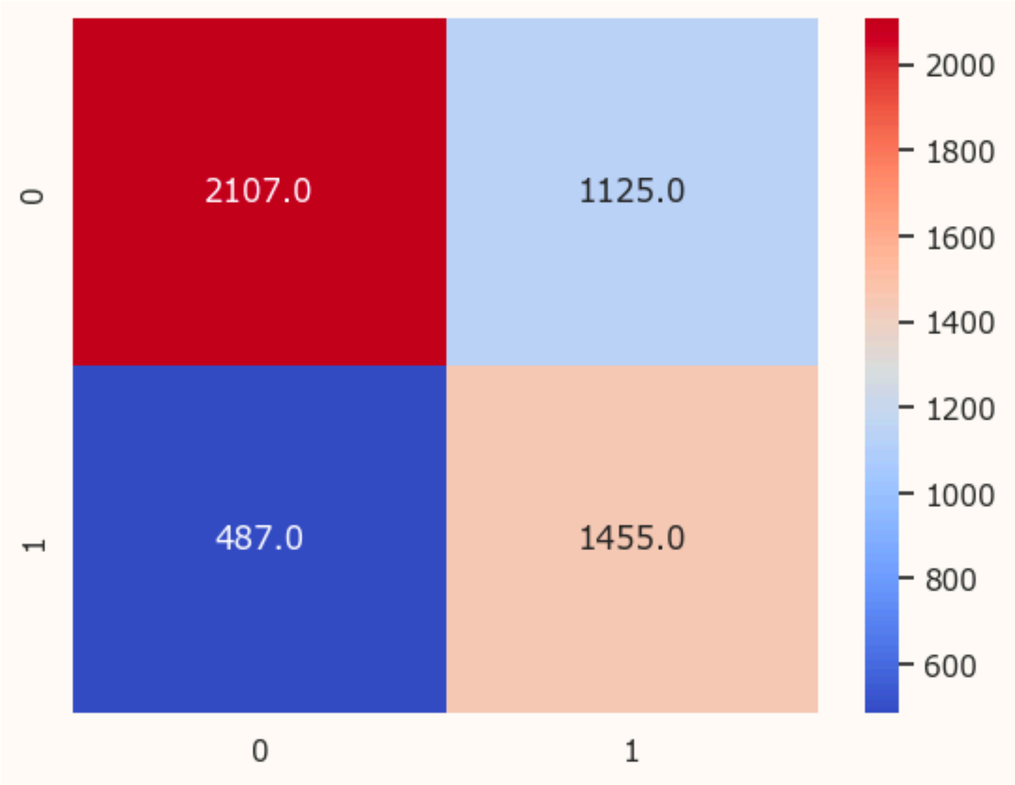
DESARROLLO DEL MODELO



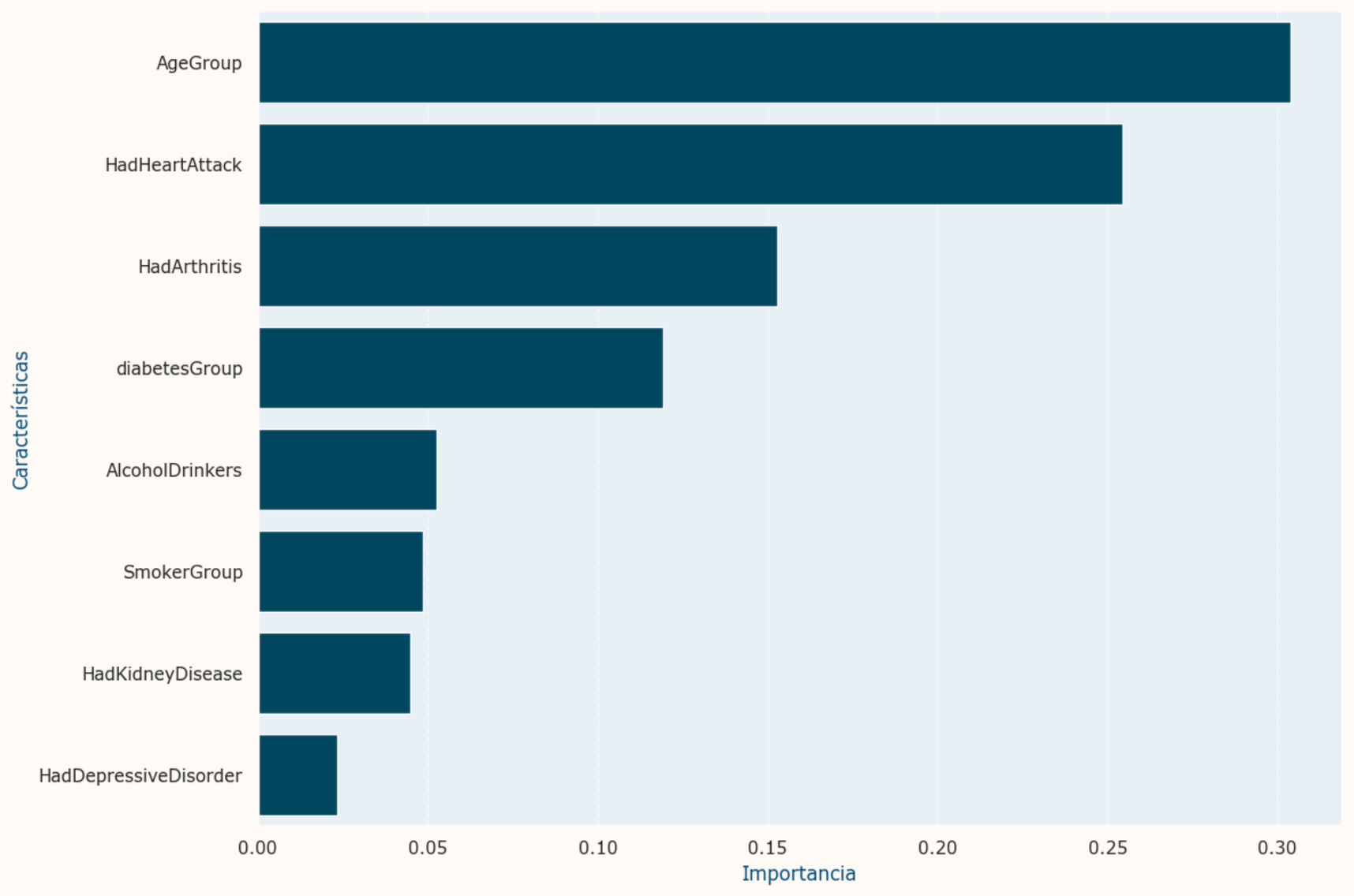
ELECCIÓN DE MODELO

RANDOM FOREST

Máximizar el recall sin perder mucha precisión



IMPORTANCIA DE CARACTERISTICAS



LIMITACIONES

- Dataset sesgado
- Mayor enfoque en busqueda de hiperparámetros

MEJORAS

- Muestra representativa
- Análisis descriptivo exhaustivo antes de modelar

GRACIAS

Thank you