

Optimizando la predicción del **Alzheimer** con modelos de aprendizaje automático.

Marta Tébar



índice

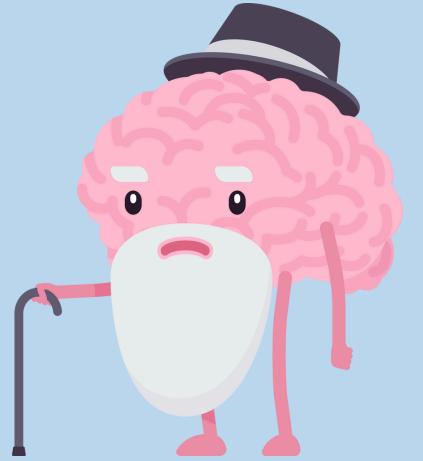
1. Introducción
2. Descripción del dataset y Mini EDA
3. Arquitectura de la solución y desarrollo del modelo
4. Resultados
5. Conclusiones

1. Introducción

Problema de negocio:

- **Gran impacto**

El **Alzheimer** afecta a un 4% de la población en España entre 75 y 79 años, y aumenta hasta un **34%** en los mayores de 85 años. También afecta a los sistemas de salud.



1. Introducción

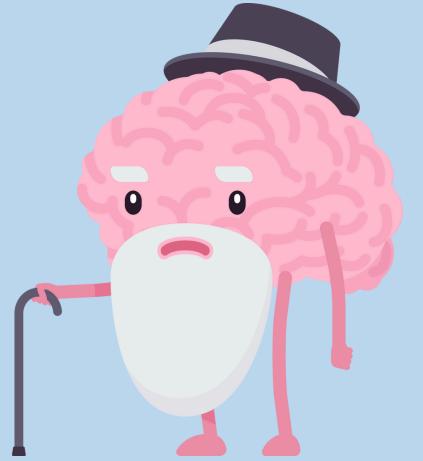
Problema de negocio:

- **Gran impacto**

El **Alzheimer** afecta a un 4% de la población en España entre 75 y 79 años, y aumenta hasta un **34%** en los mayores de 85 años.

- **Diagnóstico tardío**

Reduce opciones de tratamiento y acelera la progresión de la enfermedad.



1. Introducción

Problema de negocio:

- **Gran impacto**

El **Alzheimer** afecta a un 4% de la población en España entre 75 y 79 años, y aumenta hasta un **34%** en los mayores de 85 años.

- **Diagnóstico tardío**

Reduce opciones de tratamiento y acelera la progresión de la enfermedad.

- **Coste sanitario**

Mayor dependencia, hospitalizaciones y recursos médicos a largo plazo.



1. Introducción

Problema de negocio:

- **Gran impacto**

El **Alzheimer** afecta a un 4% de la población en España entre 75 y 79 años, y aumenta hasta un 34% en los mayores de 85 años.

- **Diagnóstico tardío**

Reduce opciones de tratamiento y acelera la progresión de la enfermedad.

- **Coste sanitario**

Mayor dependencia, hospitalizaciones y recursos médicos a largo plazo.



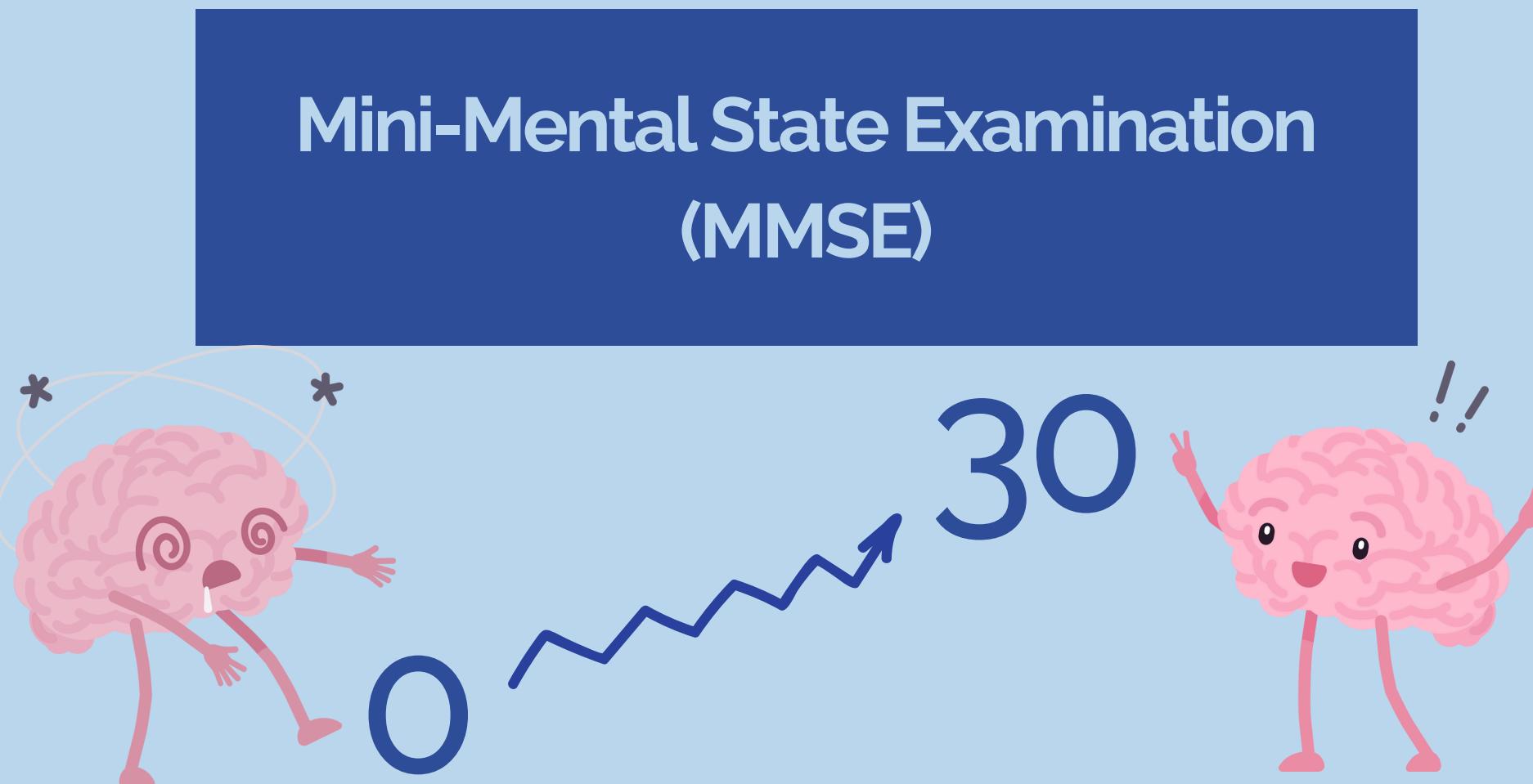
Por ello, es importante del **diagnóstico temprano**, retrasar el deterioro cognitivo, mejorar **calidad de vida** y autonomía, y **optimizar recursos** y planificación del tratamiento.

1. Introducción

Problema técnico:

- **Diagnóstico tradicional**

Basado en pruebas cognitivas e imágenes médicas. Costoso, lento y sujeto a interpretación humana.



1. Introducción

Problema técnico:

- **Diagnóstico tradicional**

Basado en pruebas cognitivas e imágenes médicas. Costoso, lento y sujeto a interpretación humana.

- **Machine Learning como solución:**

Automatización del análisis de datos clínicos, **identificación de patrones** tempranos imperceptibles para el ojo humano, mejora de la precisión y reducción de falsos diagnósticos.

2.Descripción del dataset y miniEDA

Características del dataset → Pacientes diagnosticados y no diagnosticados

Variables demográficas

Medidas clínicas

Factores de la vida diaria

Evaluaciones cognitivas y funcionales

Historial médico

Síntomas

2.Descripción del dataset y miniEDA

Características del dataset → Pacientes diagnosticados y no diagnosticados

Variables demográficas

- Edad
- Sexo
- Etnia
- Nivel de educación

2.Descripción del dataset y miniEDA

Características del dataset → Pacientes diagnosticados y no diagnosticados

Factores de la vida diaria

- Body Mass Index (BMI)
- Si fuma tabaco
- El alcohol consumido a la semana
- El ejercicio físico por semana
- La calidad de la dieta
- La calidad del sueño.

2.Descripción del dataset y miniEDA

Características del dataset → Pacientes diagnosticados y no diagnosticados

Historial médico

- Antecedentes familiares
- Enfermedades cardiovasculares
- Diabetes
- Depresión
- Traumatismo craneal
- Hipertensión.

2.Descripción del dataset y miniEDA

Características del dataset → Pacientes diagnosticados y no diagnosticados

Medidas clínicas

- Presión Arterial Media
- Colesterol Total

2.Descripción del dataset y miniEDA

Características del dataset → Pacientes diagnosticados y no diagnosticados

Evaluaciones cognitivas y funcionales

- MMSE
- Evaluación funcional
- Quejas de memoria
- Problemas de conducta
- ADL (Activities of Daily Living).

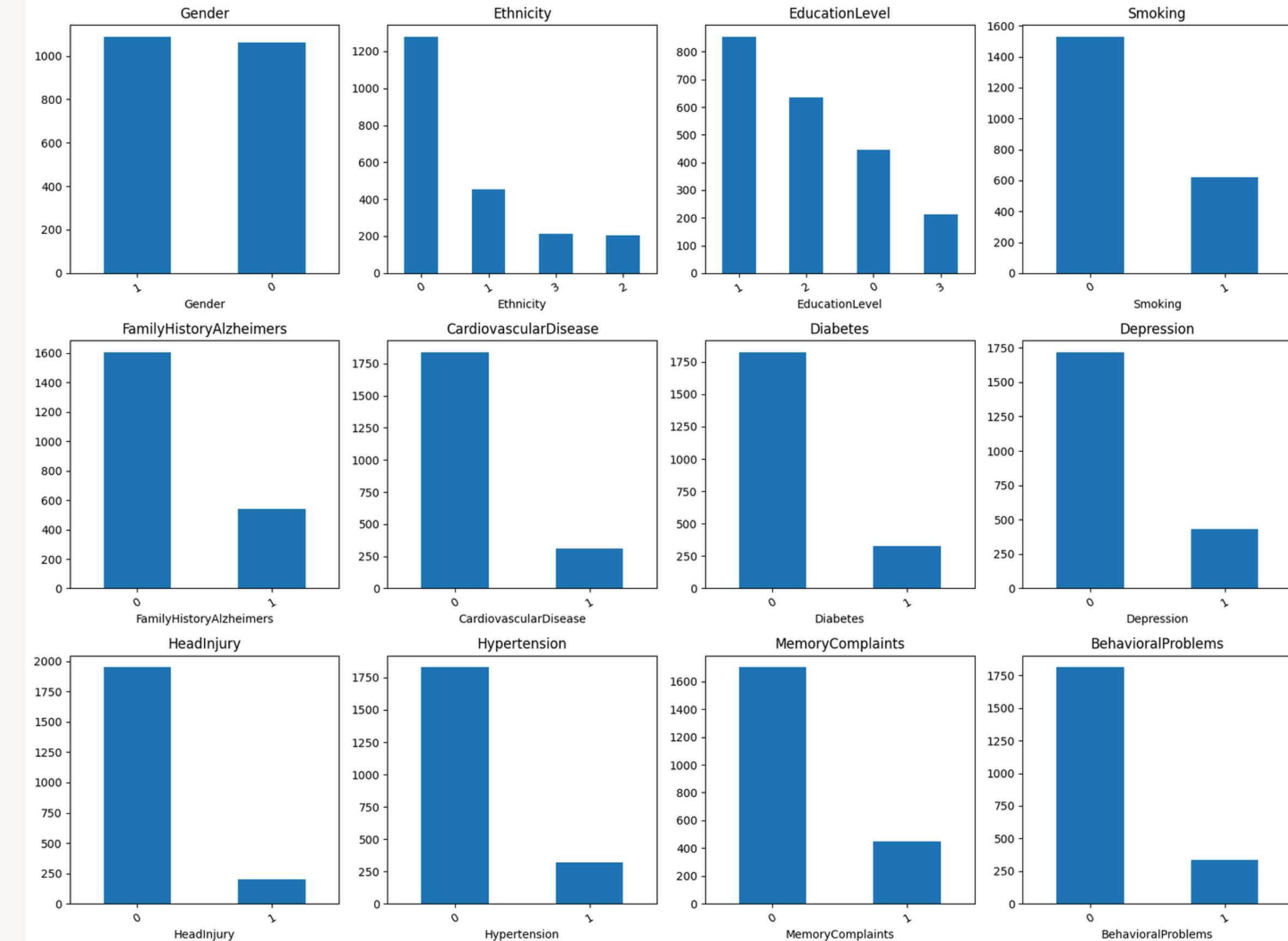
2.Descripción del dataset y miniEDA

Características del dataset → Pacientes diagnosticados y no diagnosticados

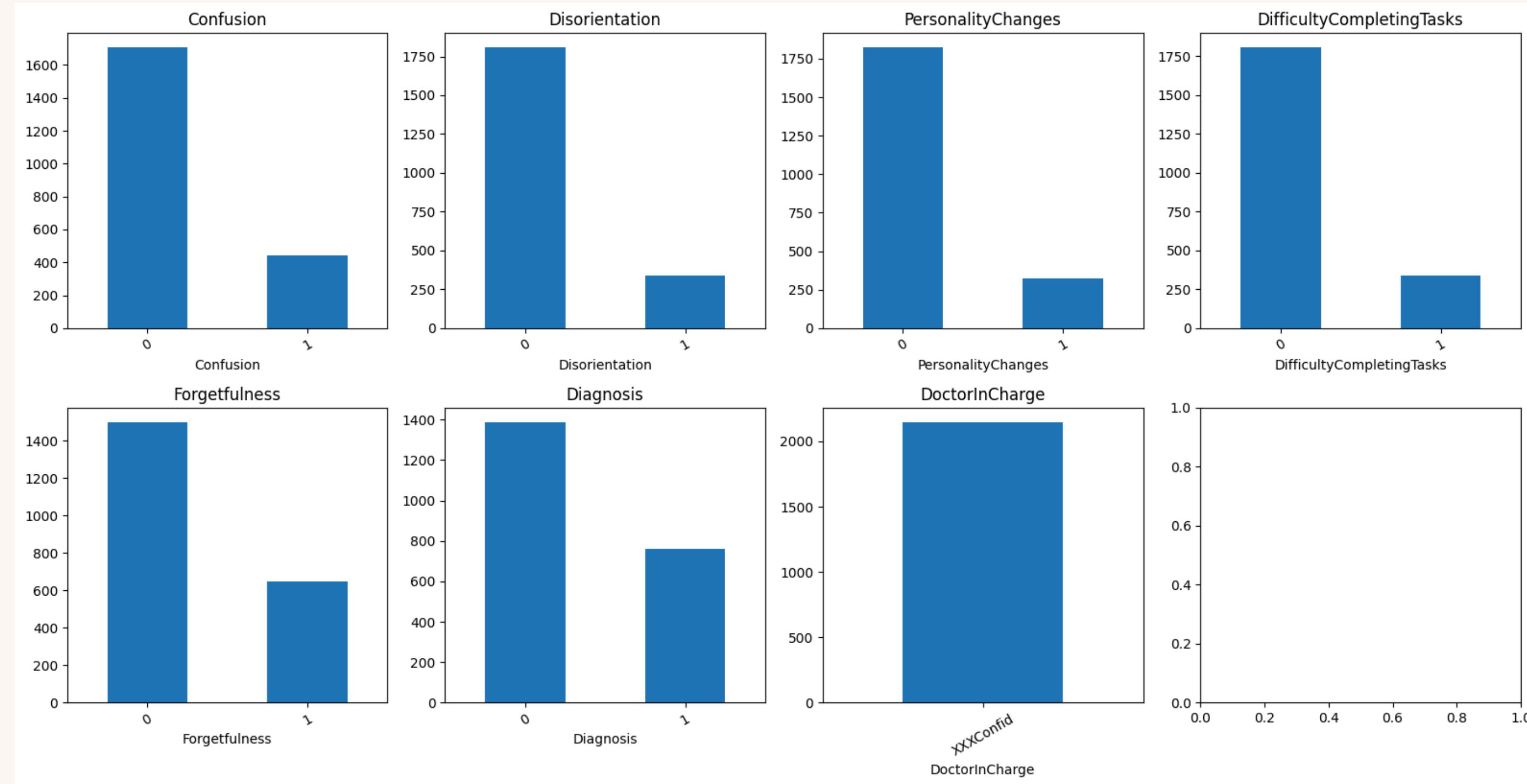
Síntomas

- Confusión
- Desorientación
- Cambios de personalidad
- Dificultad en completar tareas
- Olvido.

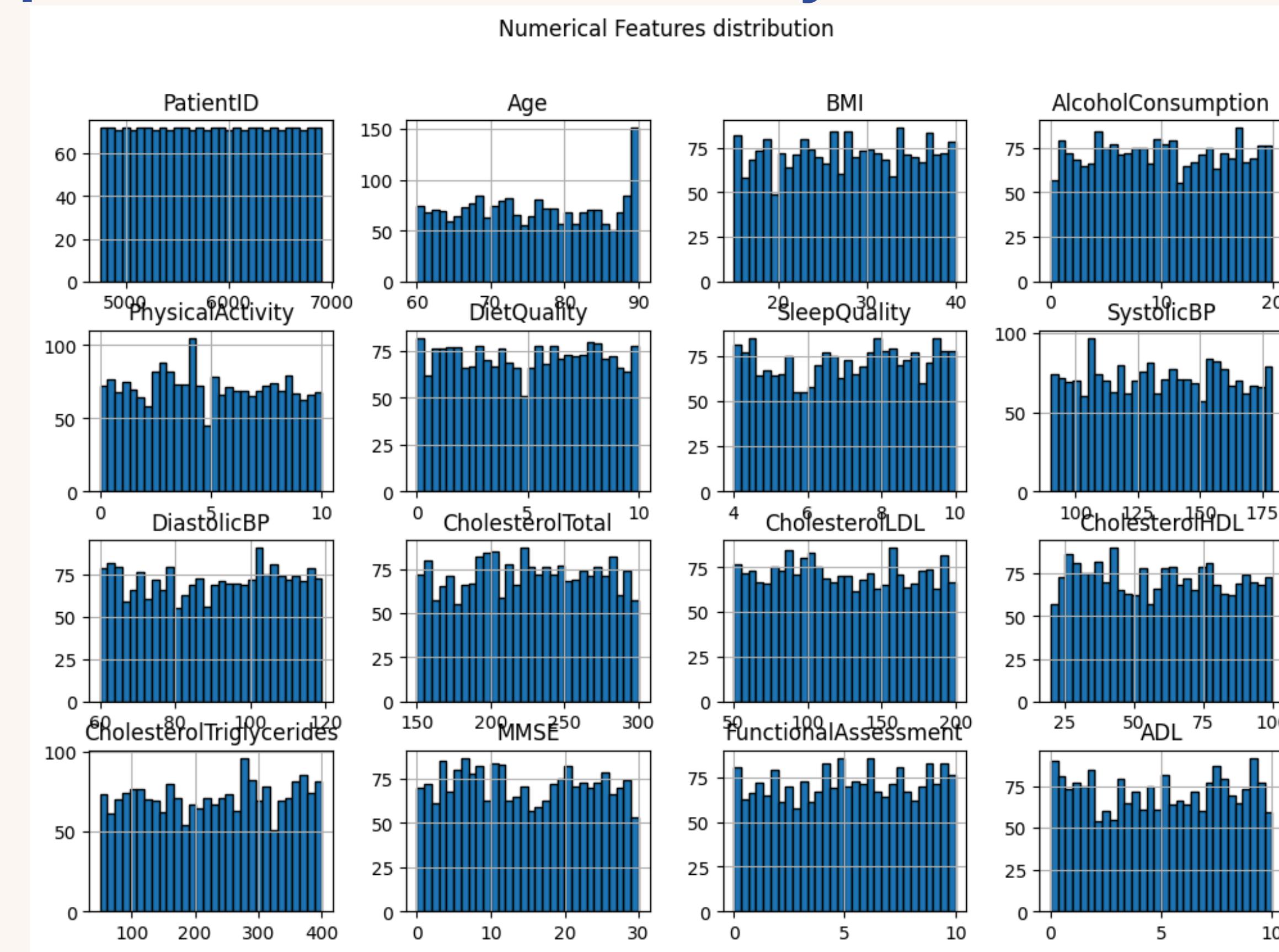
2. Descripción del dataset y miniEDA



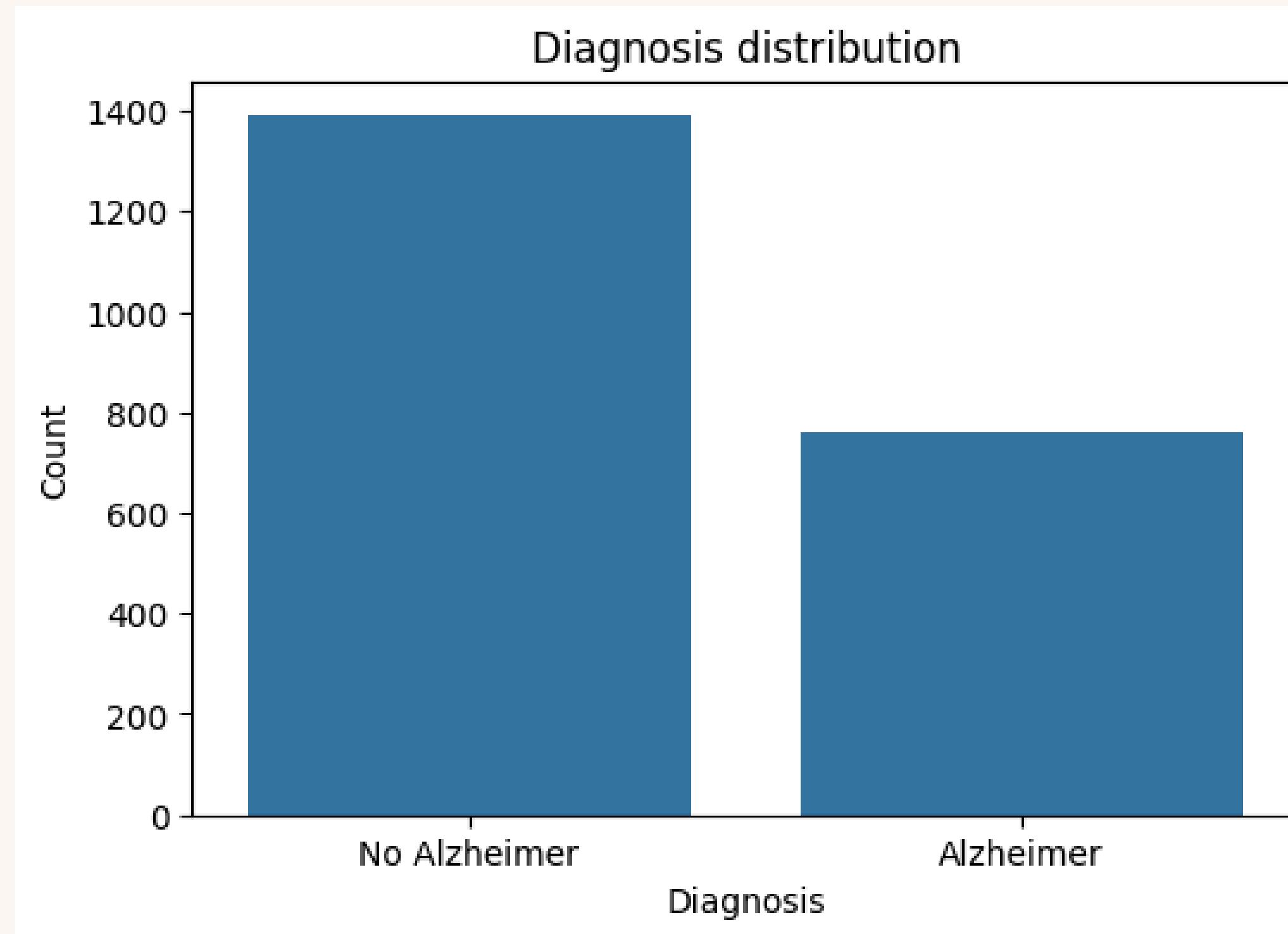
2.Descripción del dataset y miniEDA



2.Descripción del dataset y miniEDA



2.Descripción del dataset y miniEDA



- **F1-Score:** Se utiliza porque combina precisión y recall, útil en datos desequilibrados.
- **Recall:** Se centra en minimizar los falsos negativos, importante para detectar correctamente la enfermedad.

3. Arquitectura de la solución y desarrollo del modelo

Antes, una limpieza de datos y transformación de las columnas:

- Numéricas: escalado
- Categóricas: one-hot encoding

Divir datos en entrenamiento (80%) y test (20%).

01.

**Modelos
base ML**

LR, RF, SVM, KNN,
XGBoost, LightGBM

02.

**Hiperparáme-
tros**

GridSearch

03.

**Selección de
features**

RF y PCA

04.

**Deep
Learning**

Prueba

3. Arquitectura de la solución y desarrollo del modelo

01. Modelos base ML

Regresión
Logística

SVM

XGBoost

Random
Forest

KNN

LightGBM

3.Arquitectura de la solución y desarrollo del modelo

02. Hiperparámetros

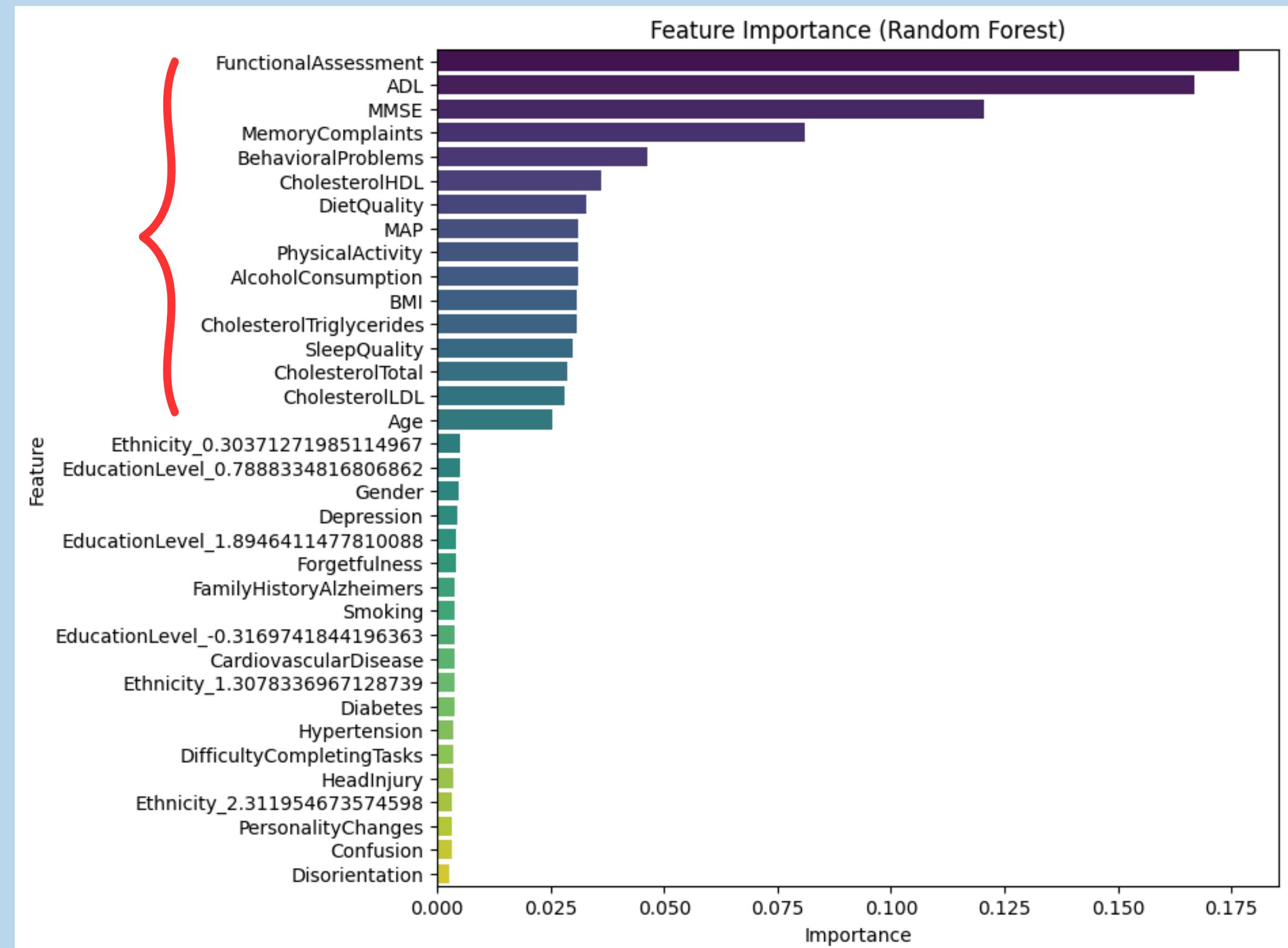
GridSearch de los tres mejores modelos.

- N_estimators
- Learning_rate
- Max_depth

3. Arquitectura de la solución y desarrollo del modelo

03. Selección de features

- **Random Forest** seleccionó estas 16 features.
- **PCA**



3. Arquitectura de la solución y desarrollo del modelo

04. Deep Learning

Red neuronal para clasificación binaria con **3 capas densas**:

- 128 y 64 neuronas con activación **ReLU**.
- **Dropout** (0.3) para evitar overfitting.
- Salida **sigmoide** (1 neurona) para probabilidad de clase.

4.Resultados

01. Modelos base ML

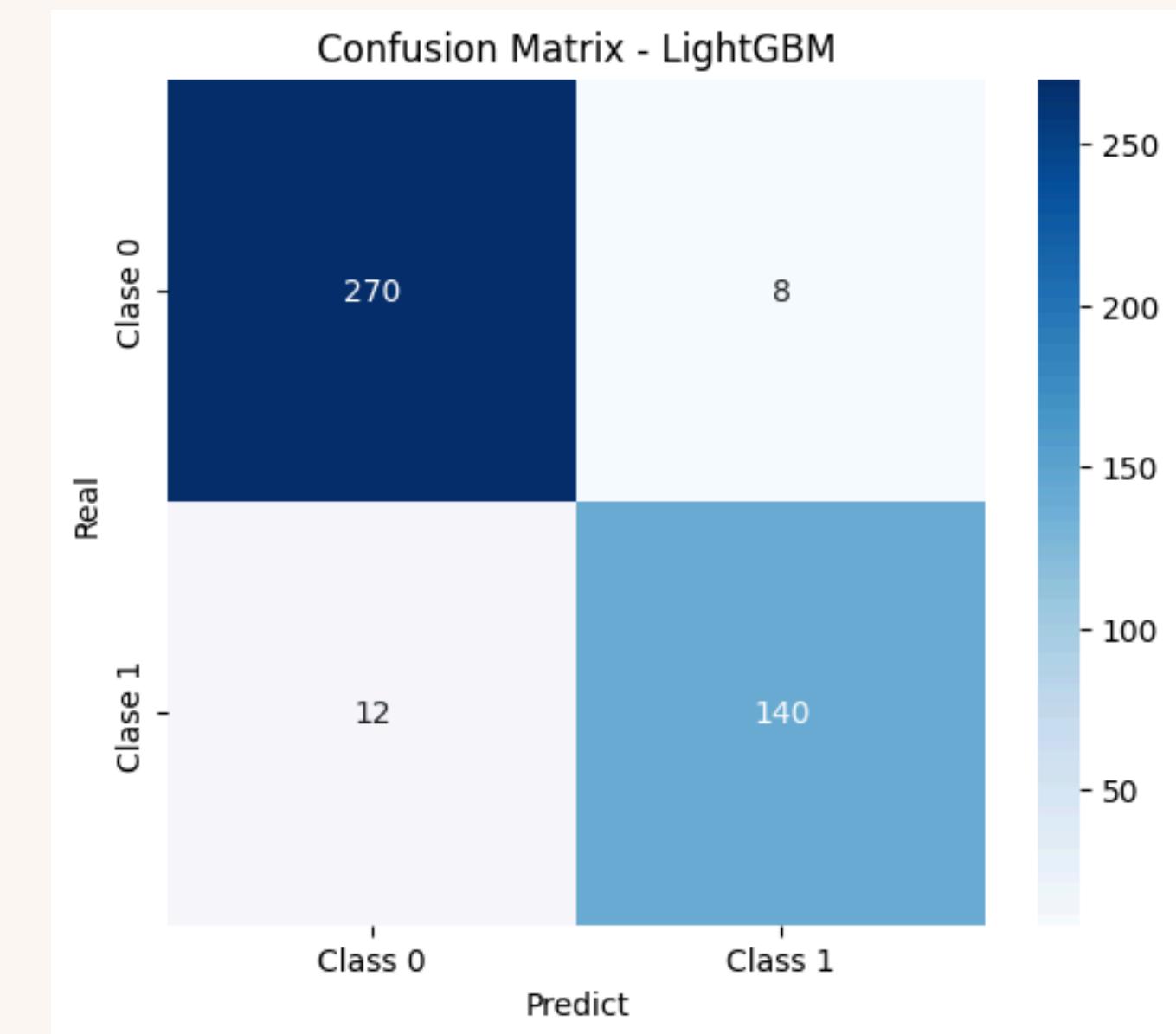
Modelo	F1-Score	Recall
LightGBM	0.929766	0.914474
XGBoost	0.926667	0.914474
Random Forest	0.907850	0.875000
SVM	0.765957	0.710526
Logistic Regression	0.754098	0.756579
KNN	0.603053	0.519737

4.Resultados

02. Hiperparámetros

LightGBM

- Los mejores hiperparámetros fueron:
'learning_rate': 0.2,
'max_depth': -1,
'n_estimators': 300
- Los resultados que se obtuvieron en test:
F1-Score: 0.9333
Recall: 0.95

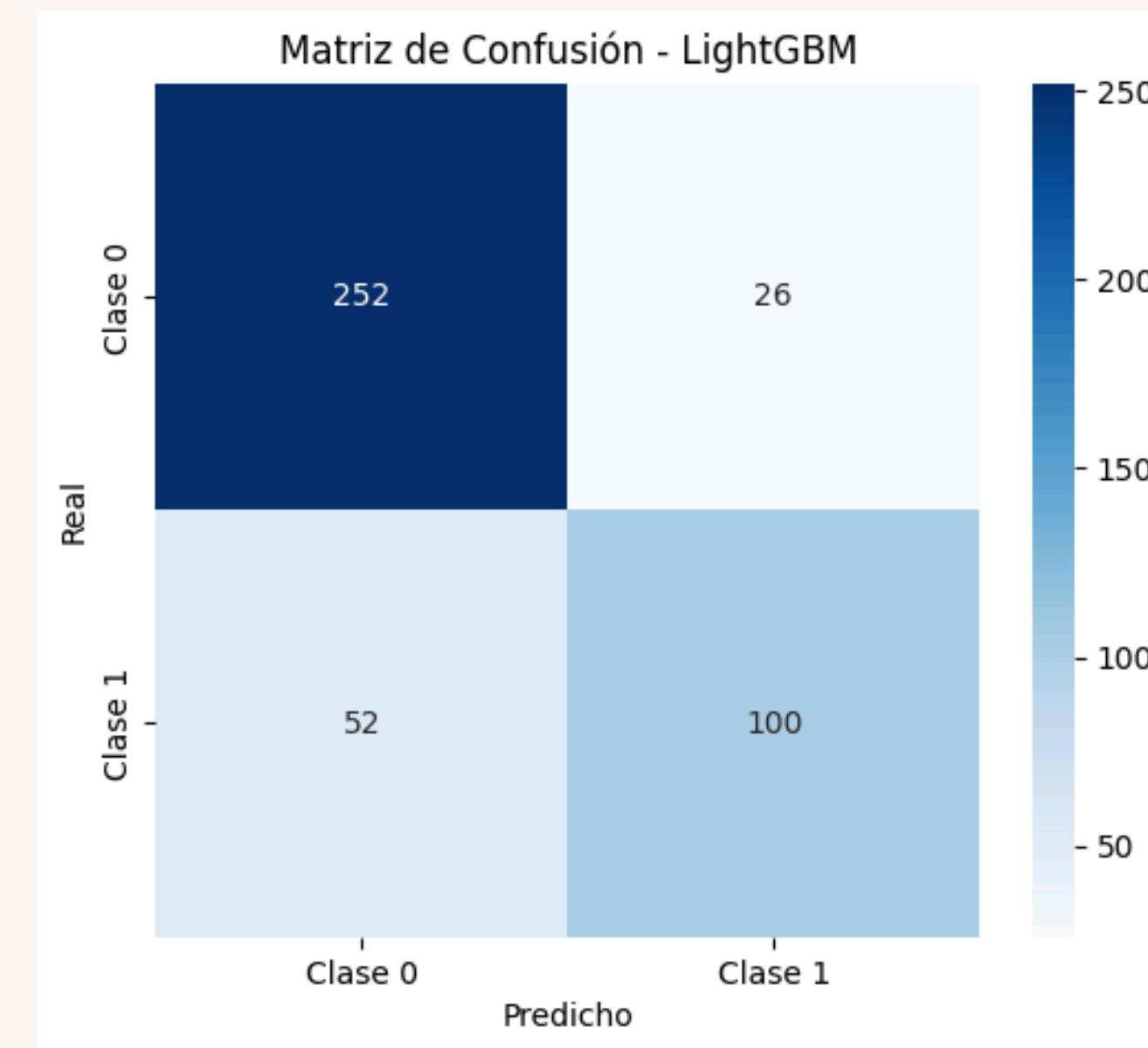


4.Resultados

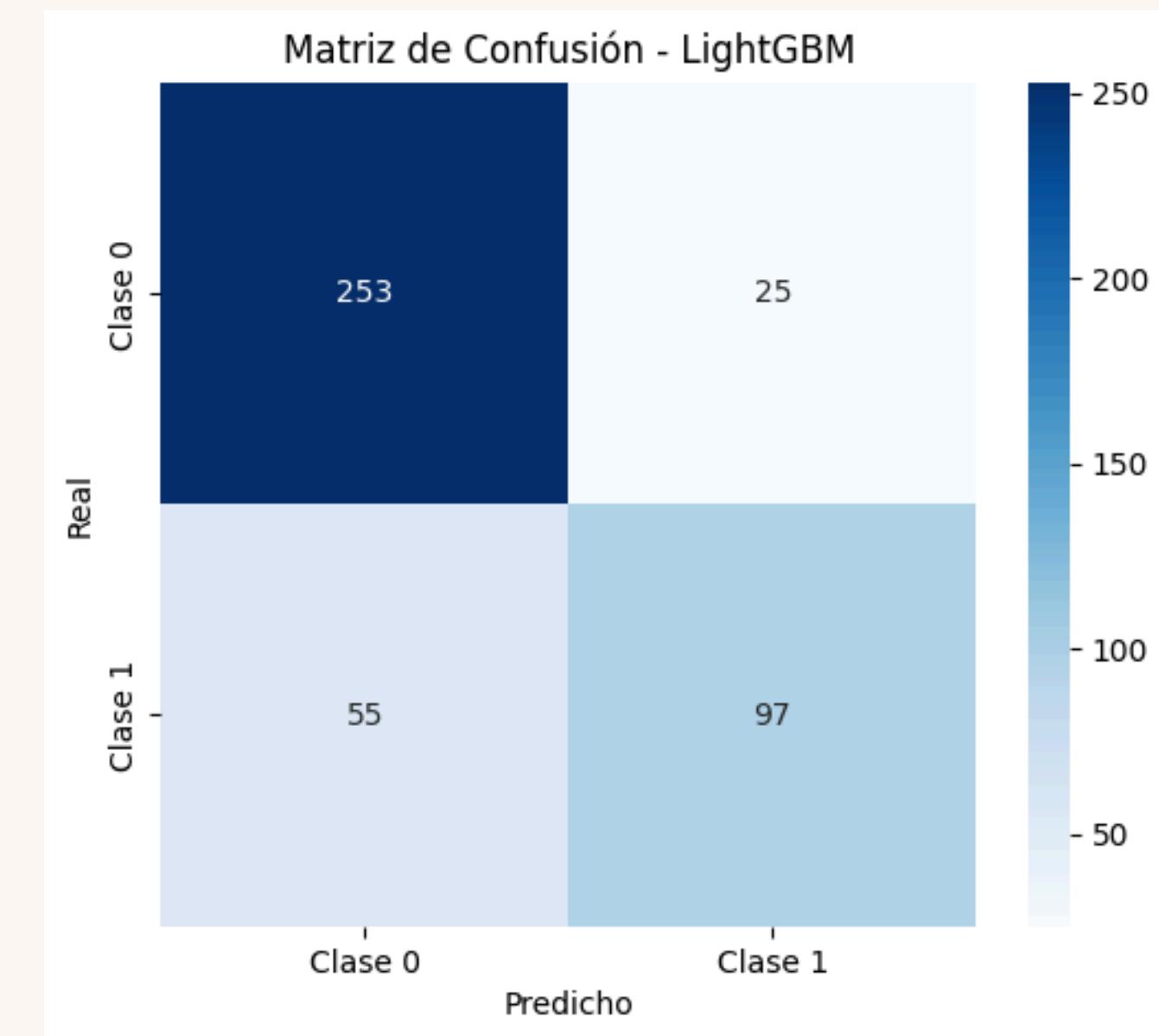
03.

Feature Selection: PCA

Resultados de **LightGBM** con el
modelo base: 0.7194



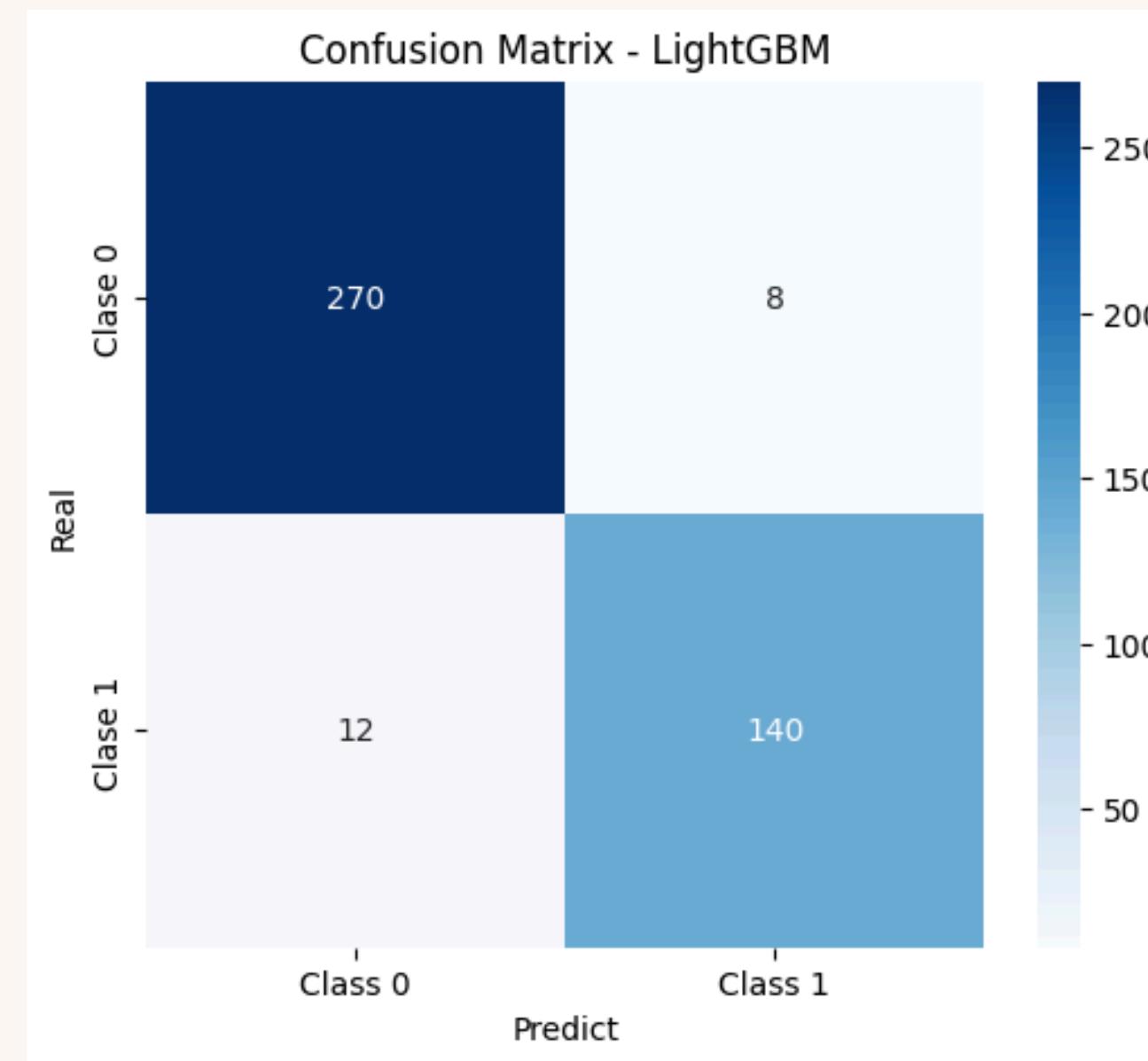
Resultados de **LightGBM** con el
hiperparámetros ajustados: 0.7080



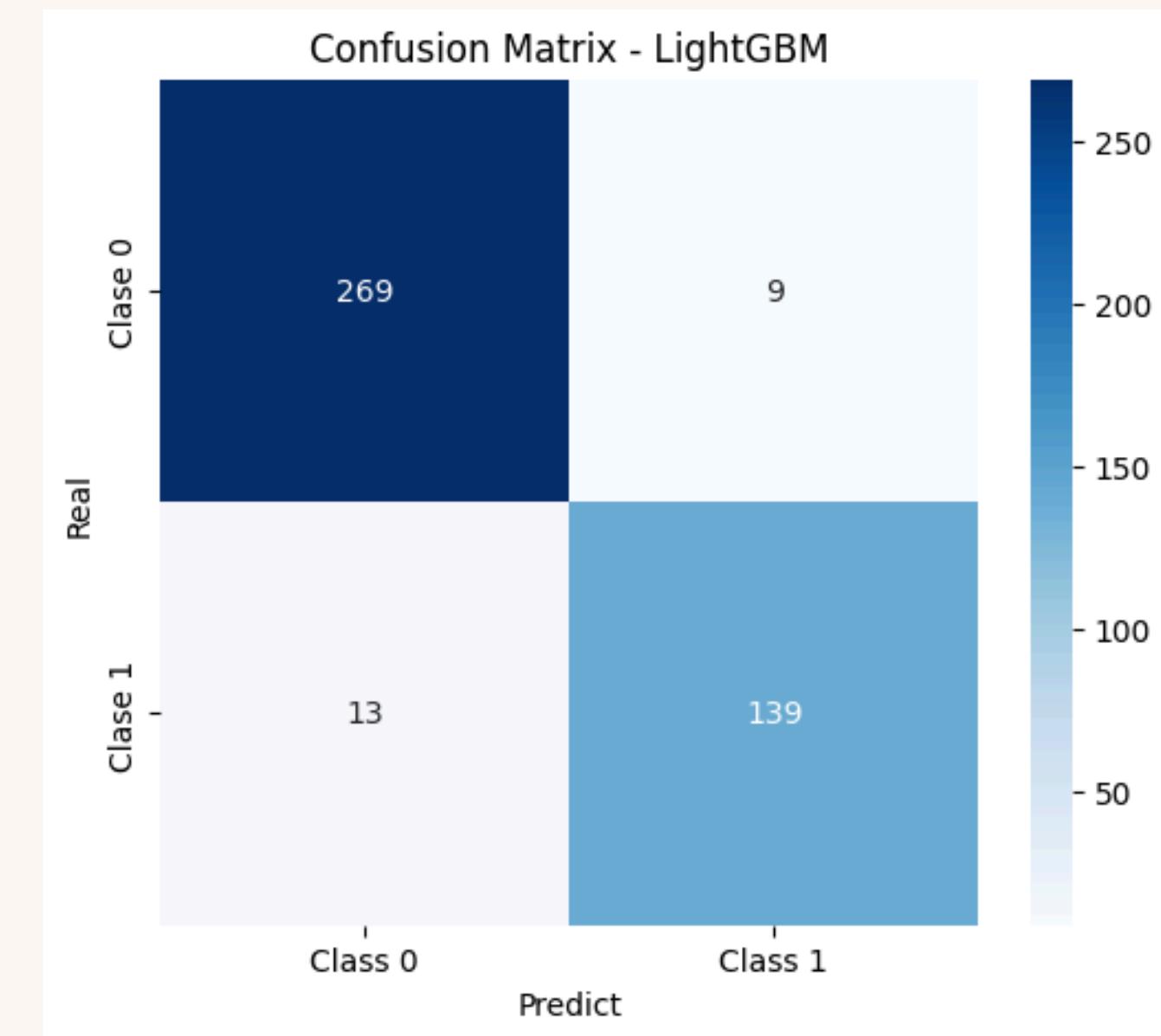
4.Resultados

03. Feature Selection: Random Forest

Resultados de **LightGBM** con el
modelo base: 0.9333



Resultados de **LightGBM** con el
hiperparámetros ajustados: 0.9266



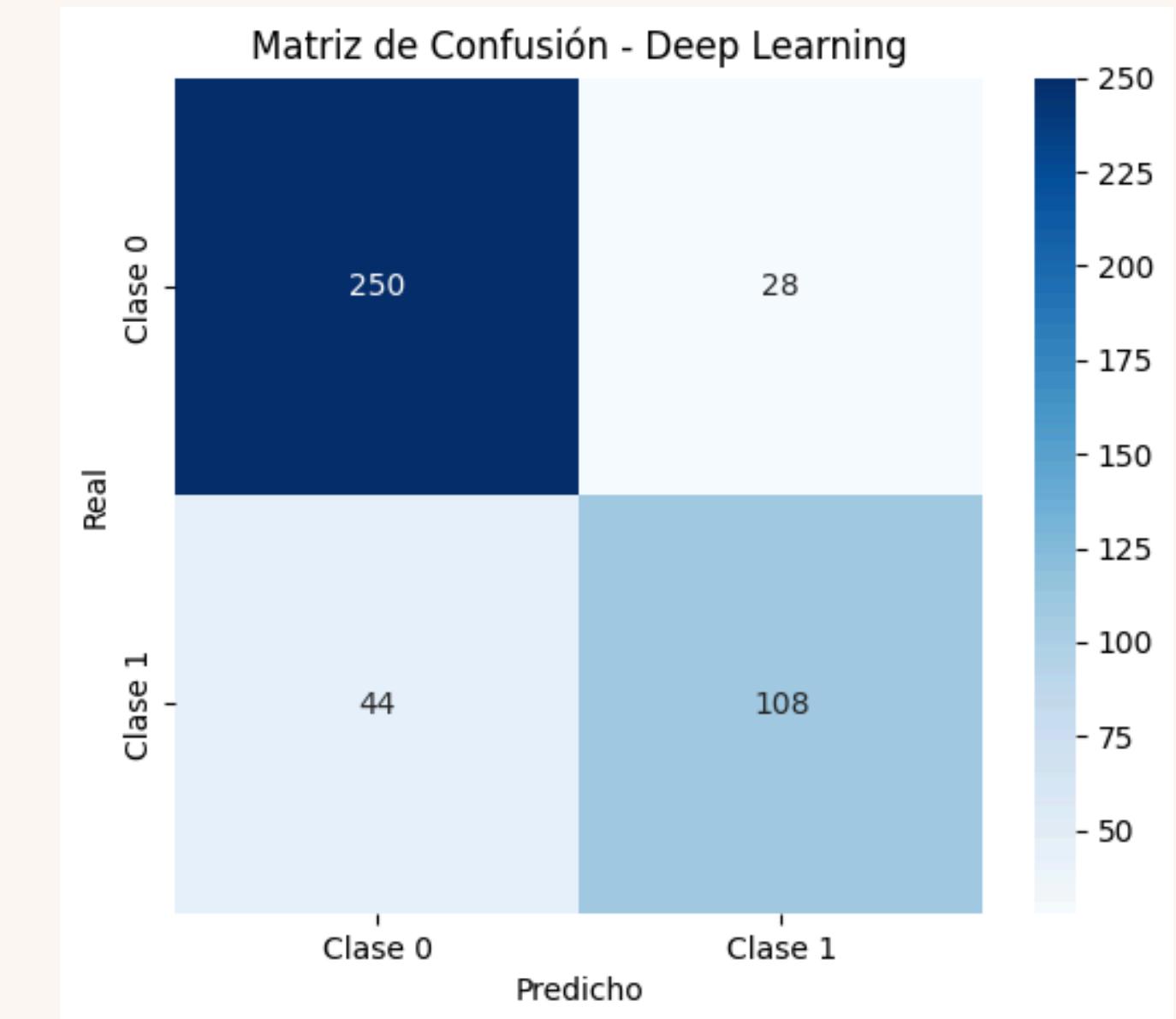
4.Resultados

04. Deep Learning

Red neuronal:

F1-Score: 0.75

- Innecesaria por la complejidad y no aporta mejores resultados que ML.



5. Conclusiones

- **Modelo base**, F1-Score de 0,9298.

5. Conclusiones

- **Modelo base**, F1-Score de 0,9298.
- **Modelo con hiperparámetros optimizados**, F1-Score de 0,9333. Indica que la optimización de los hiperparámetros ayudó.

5. Conclusiones

- **Modelo base**, F1-Score de 0,9298.
- **Modelo con hiperparámetros optimizados**, F1-Score de 0,9333. Indica que la optimización de los hiperparámetros ayudó.
- **Modelo con selección de características (base)**, F1-Score de 0,9333. El modelo base mejora hasta igualar el rendimiento del modelo con hiperparámetros optimizados. Esto sugiere que algunas variables eran irrelevantes o tenían ruido.

5. Conclusiones

- **Modelo base**, F1-Score de 0,9298.
- **Modelo con hiperparámetros optimizados**, F1-Score de 0,9333. Indica que la optimización de los hiperparámetros ayudó.
- **Modelo con selección de características (base)**, F1-Score de 0,9333. El modelo base mejora hasta igualar el rendimiento del modelo con hiperparámetros optimizados. Esto sugiere que algunas variables eran irrelevantes o tenían ruido.
- **Selección de características + hiperparámetros optimizados**, F1-Score de 0,9267. El descenso del rendimiento sugiere que los hiperparámetros se ajustaron probablemente para el conjunto completo de características y puede que ya no sean óptimos tras la reducción.

5. Conclusiones

El mejor rendimiento se obtiene con dos modelos con un **F1-Score de 0,9333**.

- Modelo con **hiperparámetros optimizados**: para asegurarse de que el modelo está bien ajustado para todo el conjunto de datos.
- Modelo **base con selección de características**: por simplicidad y eficiencia, ya que tiene menos características y el mismo rendimiento.

iMuchas
gracias!

