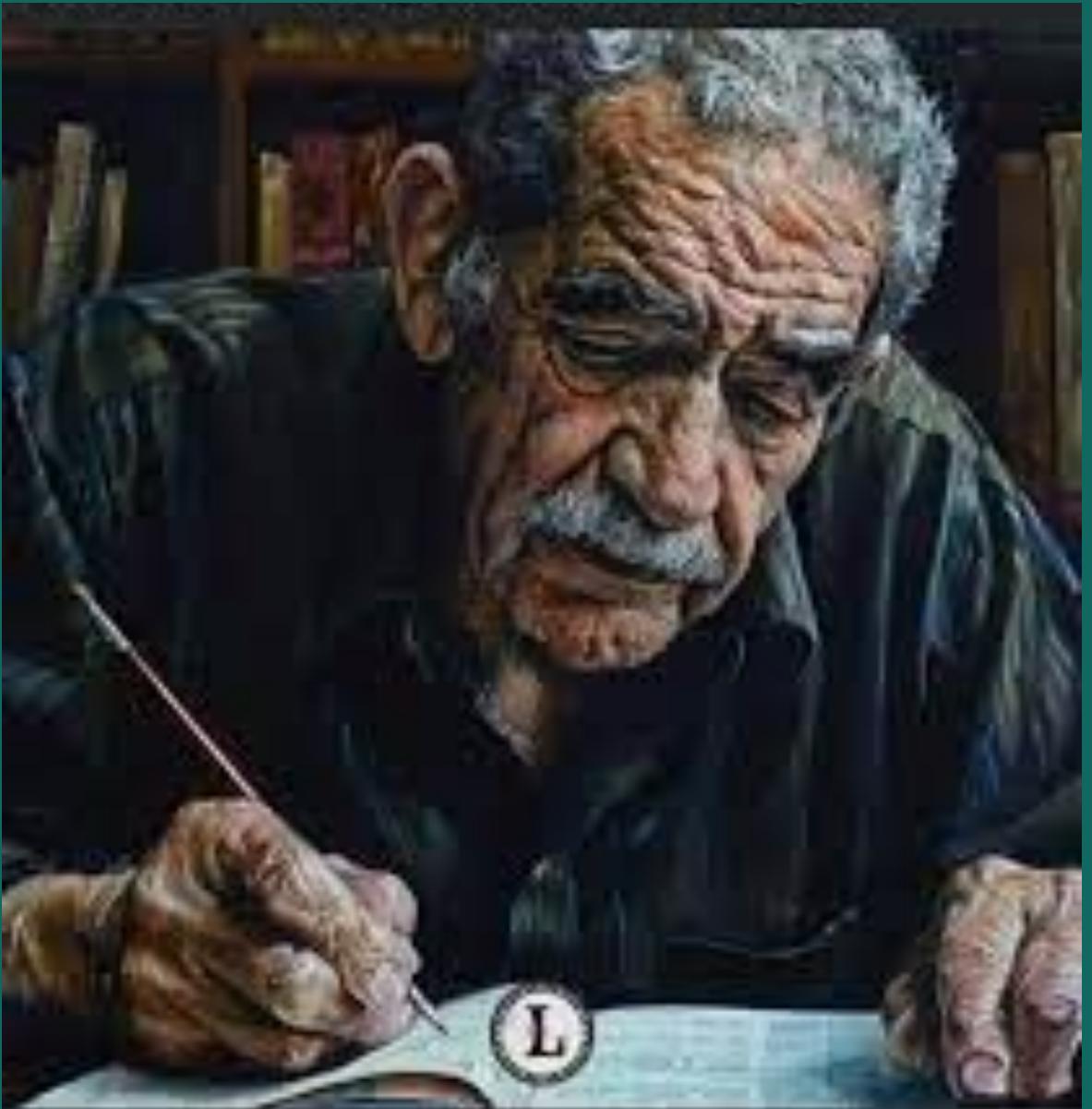


# Proyecto: Modelo predictivo de Alzheimer

Luisa Fernanda Cardenas Sierra  
Darwing Steven Sánchez Londoño  
Carlos Andrés Patiño Ibarra  
Mónica Zuluaga Quintero

*“La muerte no llega con  
la vejez sino con el  
olvido”*

*Gabriel García Marquez*





**¿Sabías que en el tiempo que dure esta presentación, al menos 3 personas habrán desarrollado Alzheimer sin siquiera saberlo?**

**Más del 50% de los casos de Alzheimer se diagnostican demasiado tarde**

**¿Qué sentirías si la persona que mas amas te olvida?**

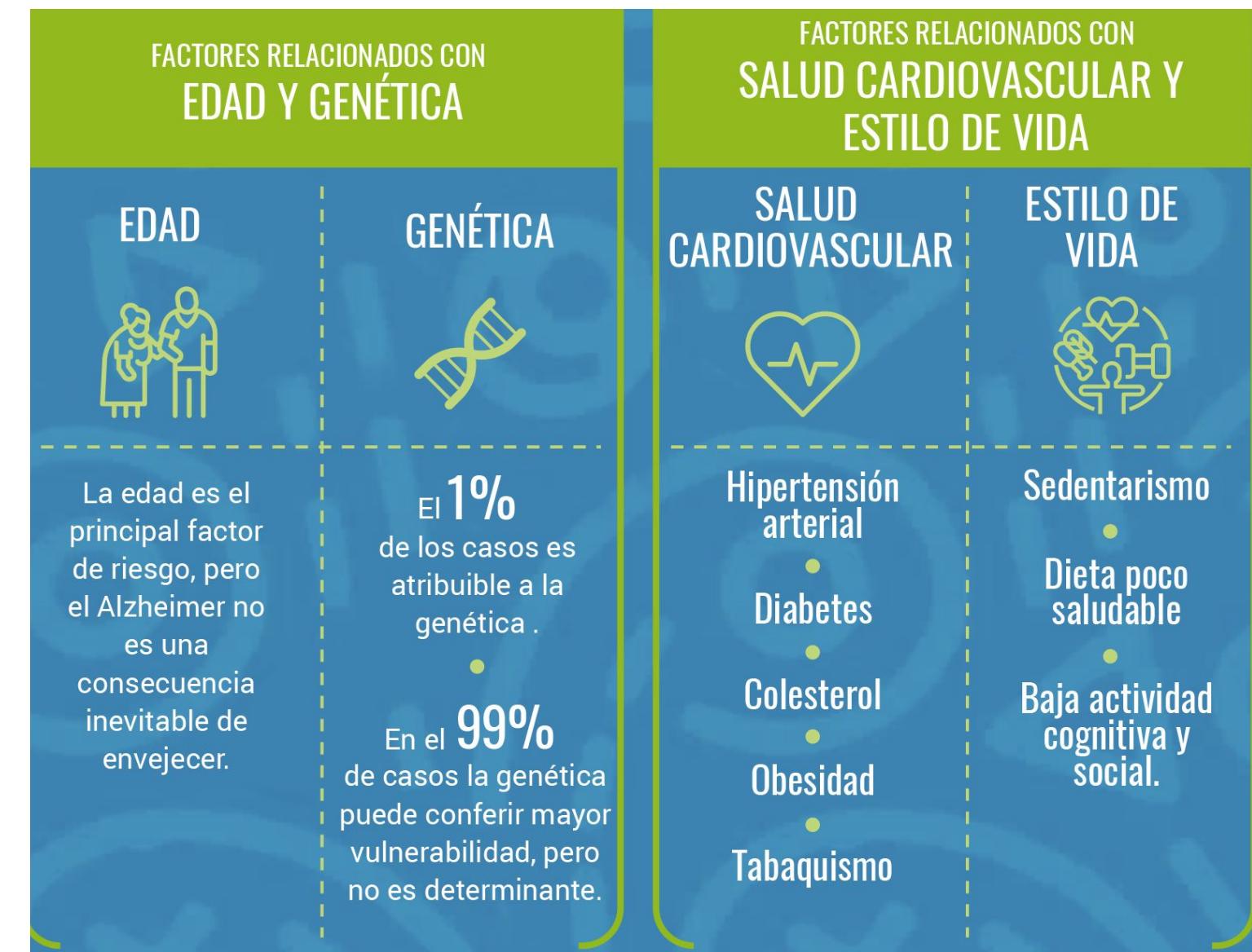
# Enfermedad de Alzheimer

Enfermedad neurodegenerativa irreversible

50-70% de todas las demencias

En América el número de pacientes con Alzheimer incrementará de 5.8 millones a 13.8 millones para 2050

Factores ambientales y comportamentales tienen un rol clave en la progresión



# Problema: diagnóstico tardío

- El Alzheimer no tiene cura, pero un diagnóstico temprano puede ralentizar la enfermedad y mejorar la calidad de vida.
- La detección actual es costosa, tardía e inaccesible para muchas personas.
- Los sistemas de salud gastan miles de millones tratando a pacientes sin herramientas de predicción eficaces.



# Por qué un modelo predictivo con IA ?



“

*La vida no es la que uno vivió,  
sino la que uno recuerda y  
cómo la recuerda para contarla.*

*GGM”*

# Objetivos



## Diagnóstico temprano

Análisis de características clínicas y sociodemográficas predictivas



## Predicción del riesgo

Personalizar la prevención y el tratamiento para cada paciente.



## Apoyo clínico y al sistema de salud

Agilizar diagnósticos, reducir costos y mejorar la cobertura en diferentes niveles de atención.

# Metodología



# Fuente de datos

# Características de los datos

DataSet contiene 74.283 registros de 20 países

Fuente: <https://www.kaggle.com/datasets/ankushpanday1/alzheimers-prediction-dataset-global>

1	Country	Age	Gender	Education	BMI	Physical Ac	Smoking	Alcohol	Diabet	Hyperte	Cholesterol	Family H	Cognitive Te	Depression	Sleep Qt	Dietary Hab	Air Pollu	Employme	Marital S	Genetic	Social	E Income	I Stress	Urban v	Alzheimer Dia		
2	Russia	59	Male		9	29.6	Low	Former	Occasional	No	No	High	No		37	Medium	Poor	Unhealthy	Low	Retired	Married	No	High	High	Low	Urban	No
3	Sweden	90	Male		2	25.2	Low	Former	Never	No	Yes	Normal	No		65	Medium	Average	Healthy	High	Retired	Widowed	No	High	Medium	High	Rural	No
4	UK	76	Female		9	33.6	High	Former	Regularly	No	No	Normal	Yes		81	High	Good	Healthy	Low	Employed	Widowed	No	Medium	High	High	Urban	No
5	USA	57	Female		0	30.7	Low	Current	Regularly	No	Yes	High	Yes		58	High	Good	Unhealthy	Low	Employed	Single	No	High	Medium	Medium	Rural	No
6	Norway	92	Female		16	27.7	Low	Never	Regularly	No	No	Normal	No		85	Medium	Average	Average	High	Retired	Widowed	No	High	Medium	High	Urban	No
7	Norway	79	Female		18	26.1	High	Former	Occasional	No	Yes	Normal	Yes		59	High	Average	Healthy	Medium	Employed	Single	No	Low	Low	Medium	Urban	No
8	India	63	Female		2	31.4	Medium	Never	Never	No	Yes	High	Yes		63	Low	Average	Unhealthy	Low	Employed	Single	Yes	Medium	High	Low	Urban	No
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....		

25 en total

## Features

- Datos demográficos, clínicos, genéticos, estilo de vida

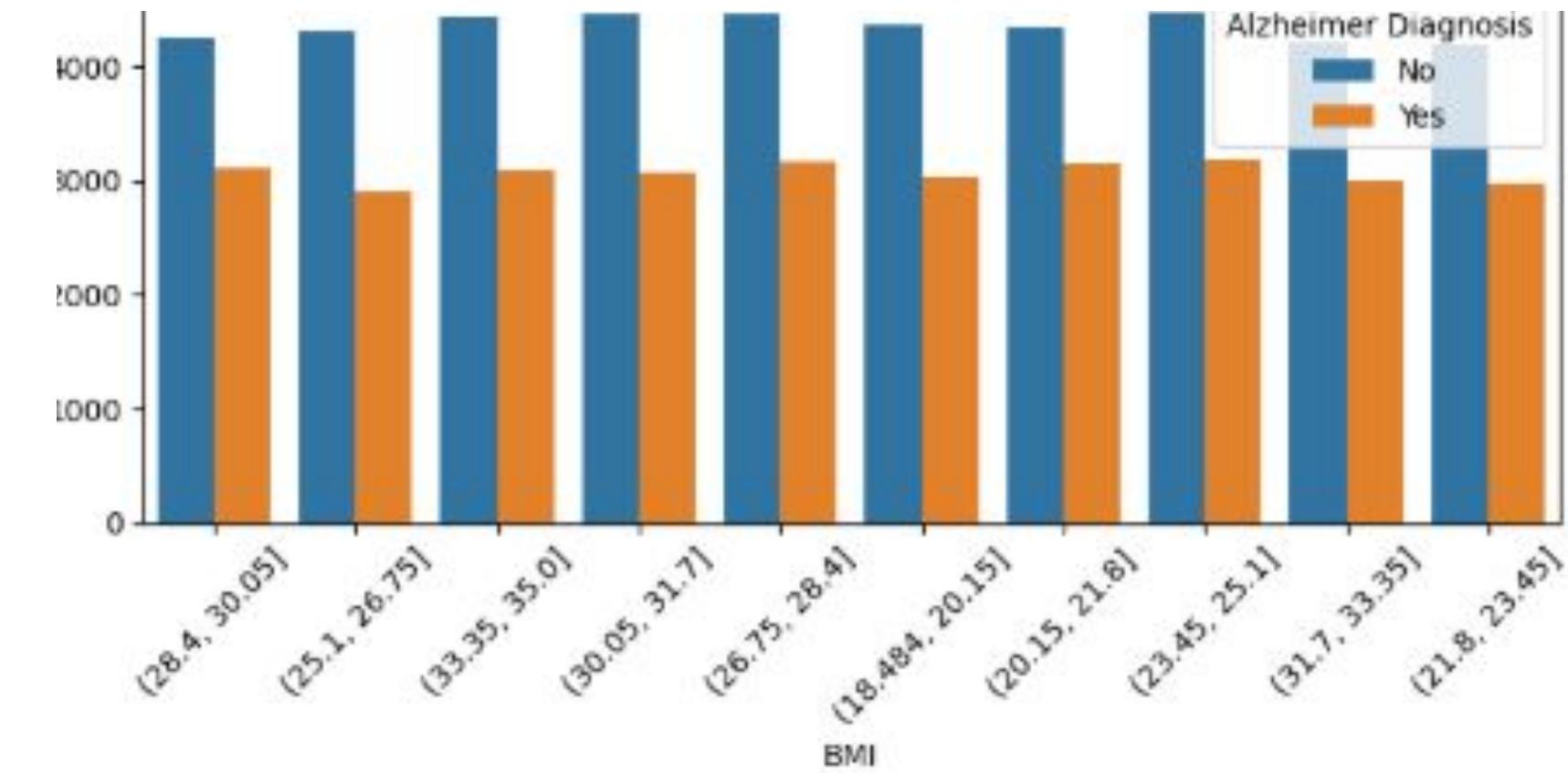
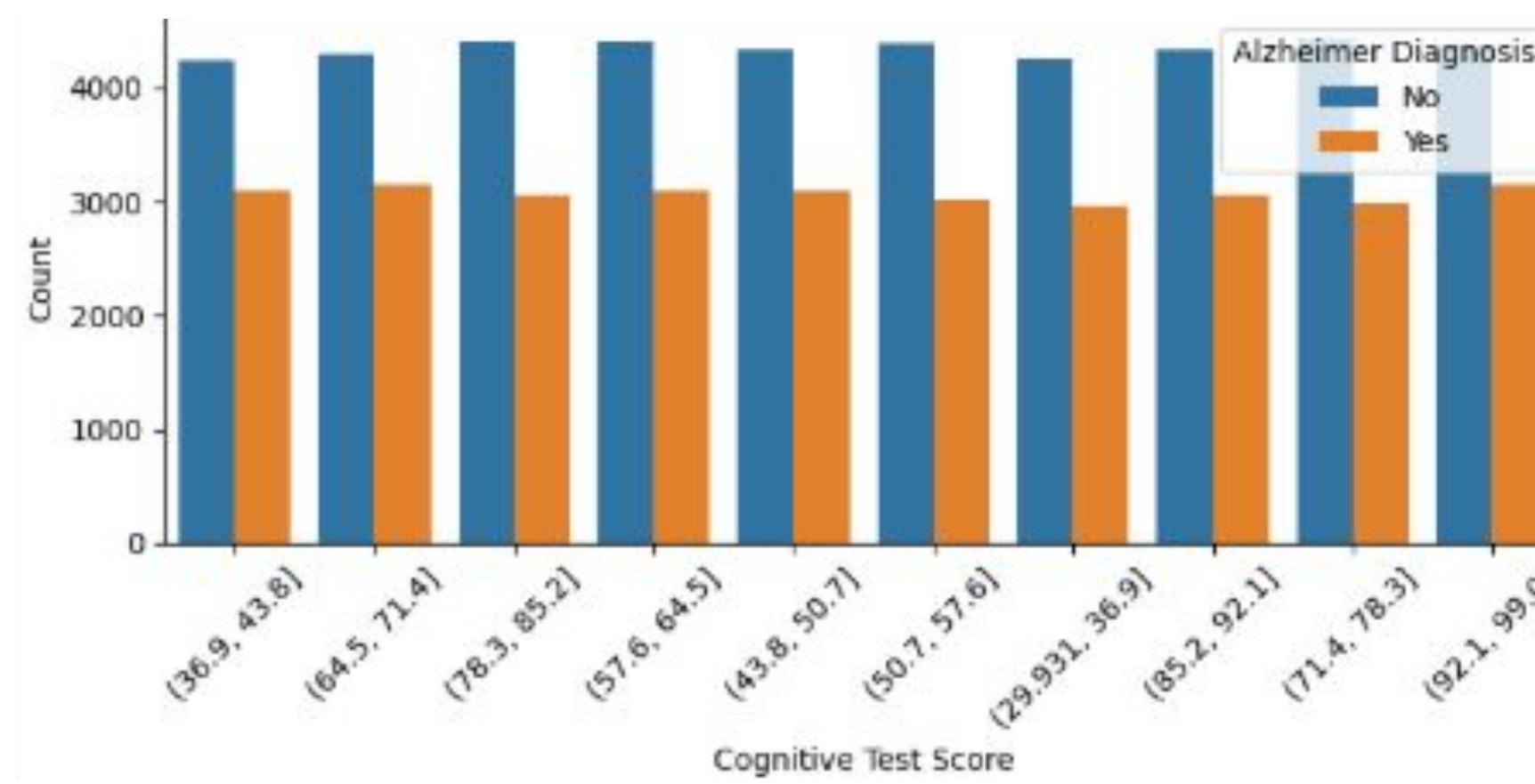
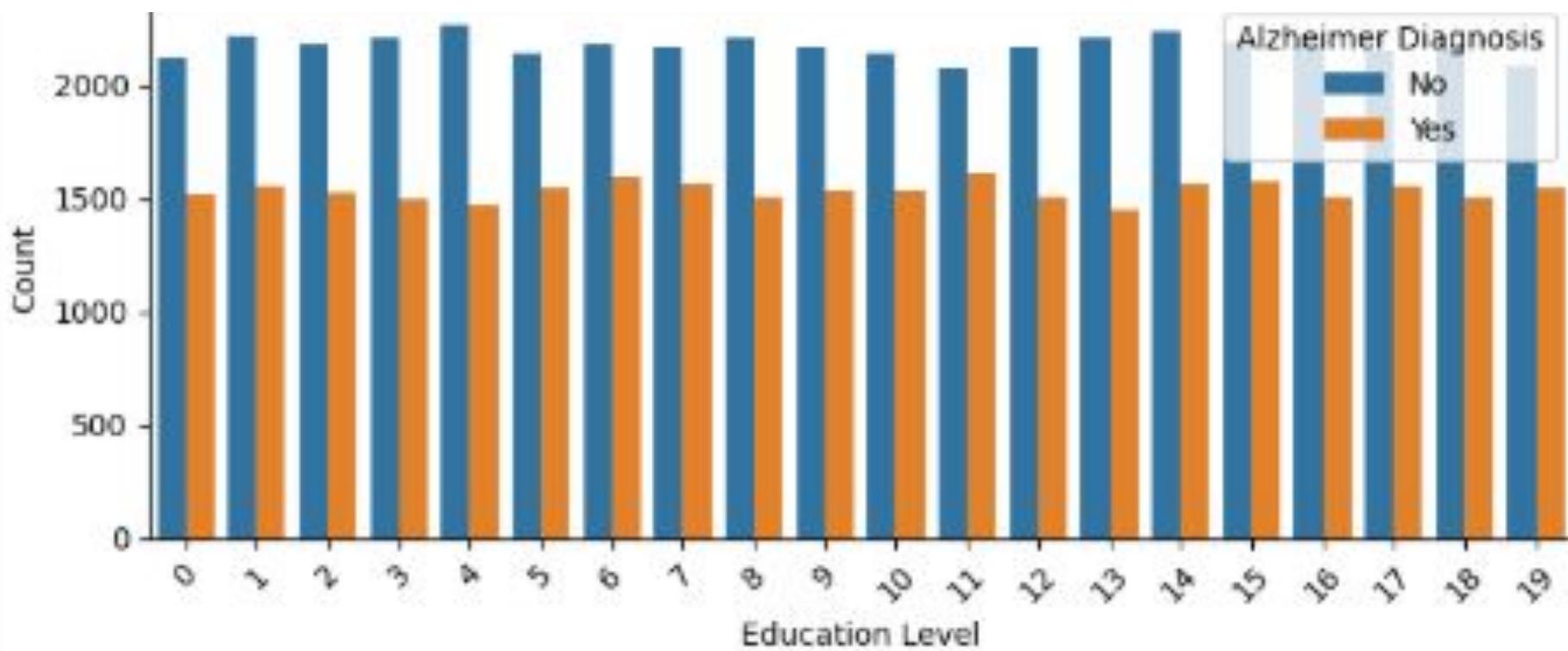
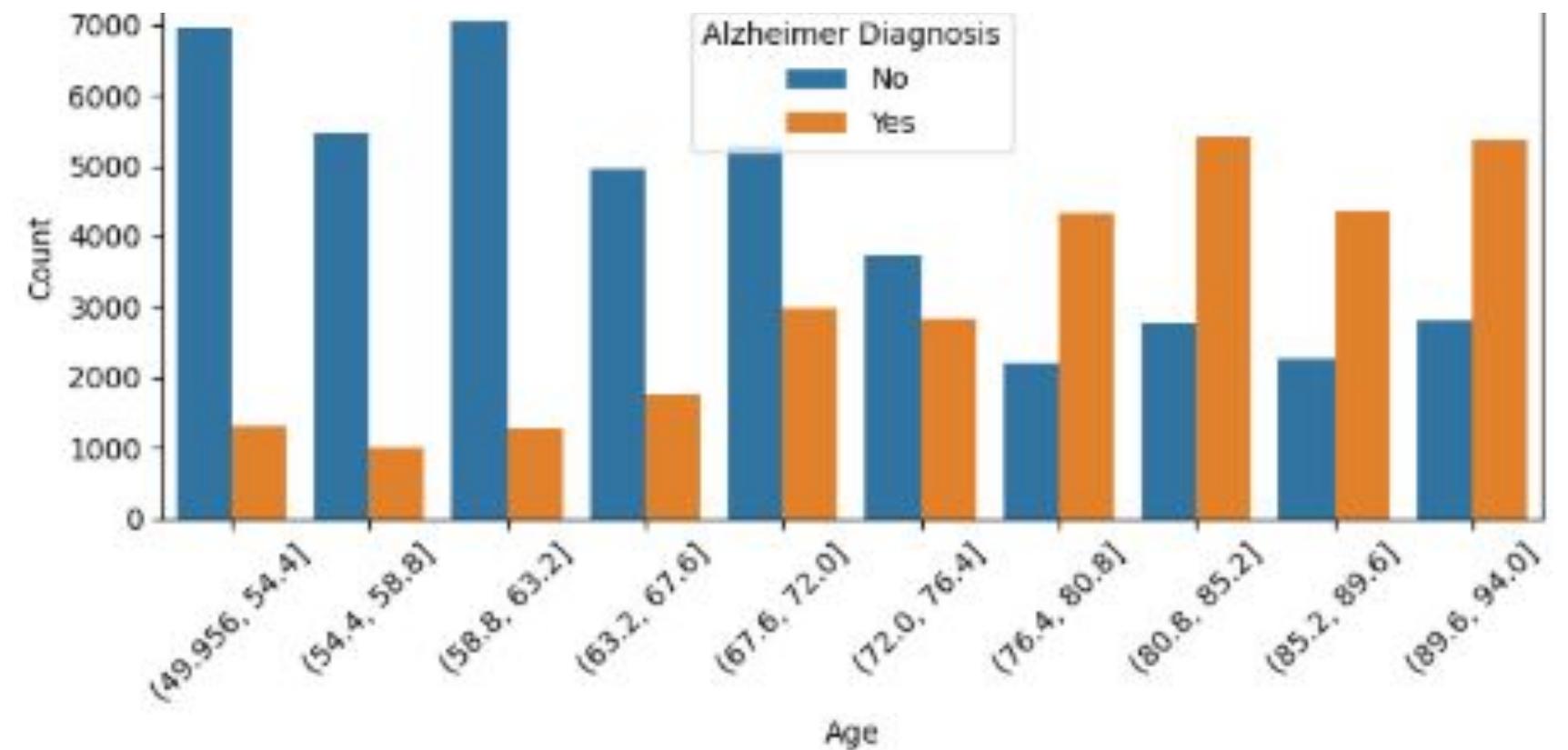
## Instancias

- 74183--- Entrenamiento/prueba (80/20%) 100— evaluación datos
- 21 categóricas, 4 numéricas: 3 enteros, 1 decimal

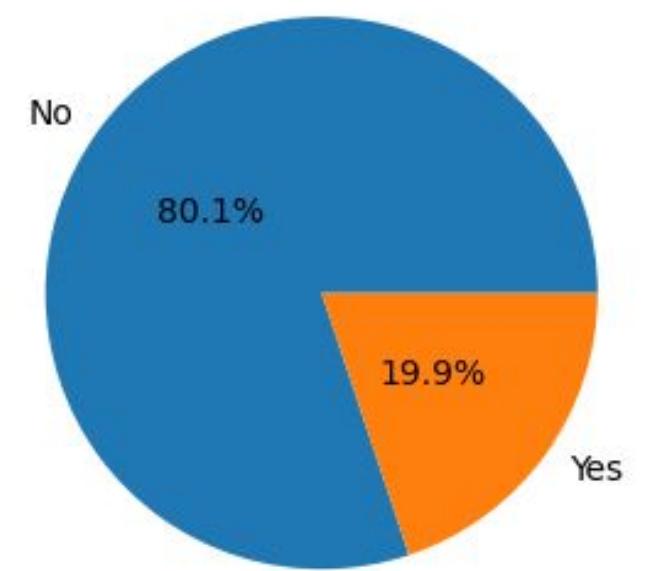
Licencia MIT



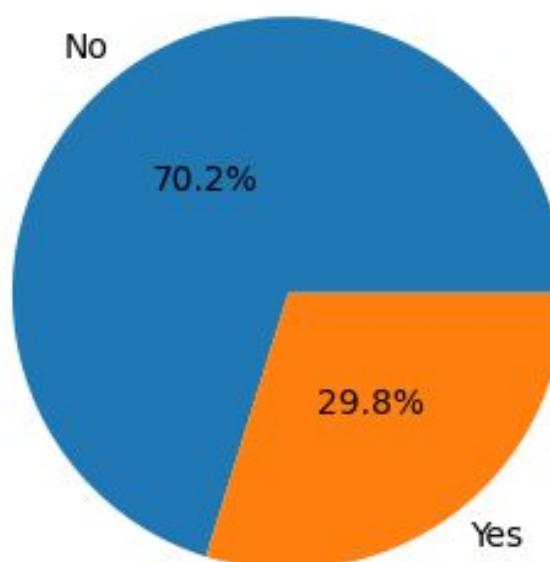
# Análisis exploratorio de datos



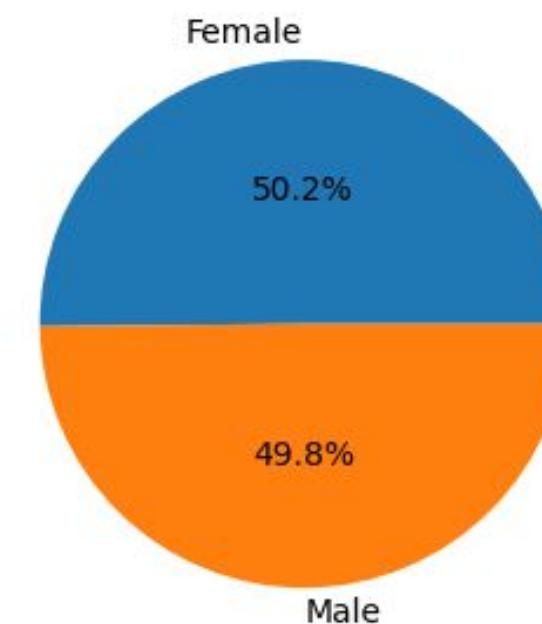
Diabetes vs Alzheimer Diagnosis



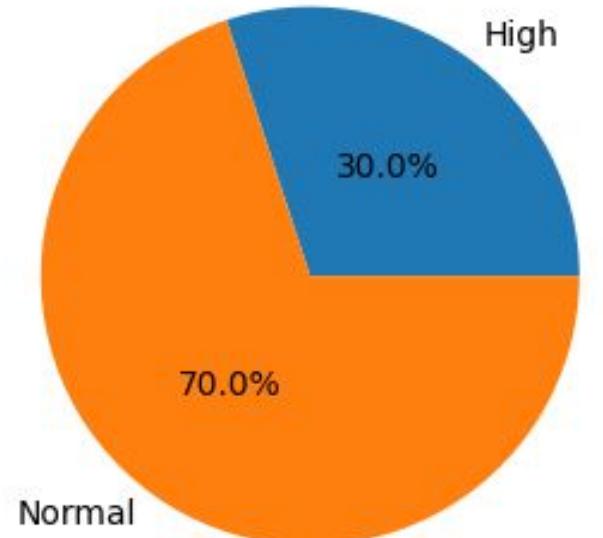
Hypertension vs Alzheimer Diagnosis



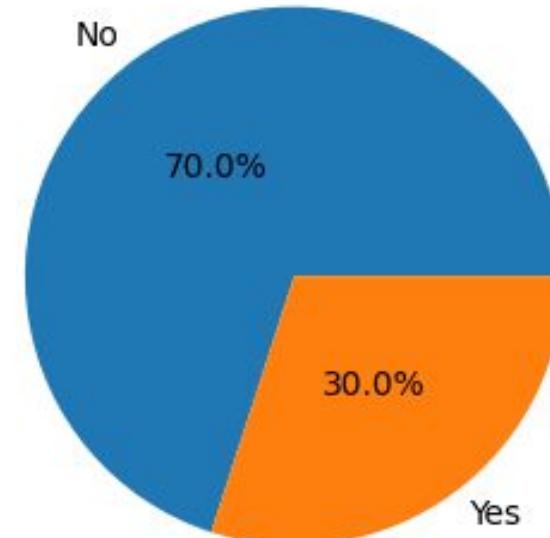
Gender vs Alzheimer Diagnosis



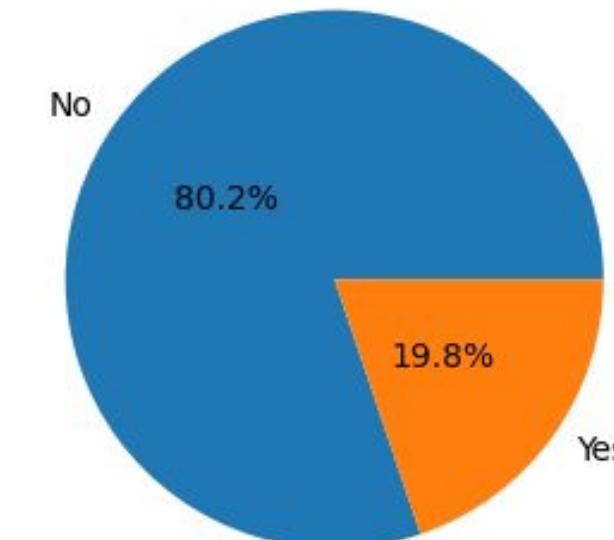
Cholesterol Level vs Alzheimer Diagnosis



Family History of Alzheimers vs Alzheimer Diagnosis



Genetic Risk Factor (APOE-E4) vs Alzheimer Diagnosis





# Preprocesamiento de datos

# Preparación de los datos

## Imputación

No valores nulos, ni duplicados

## Codificación

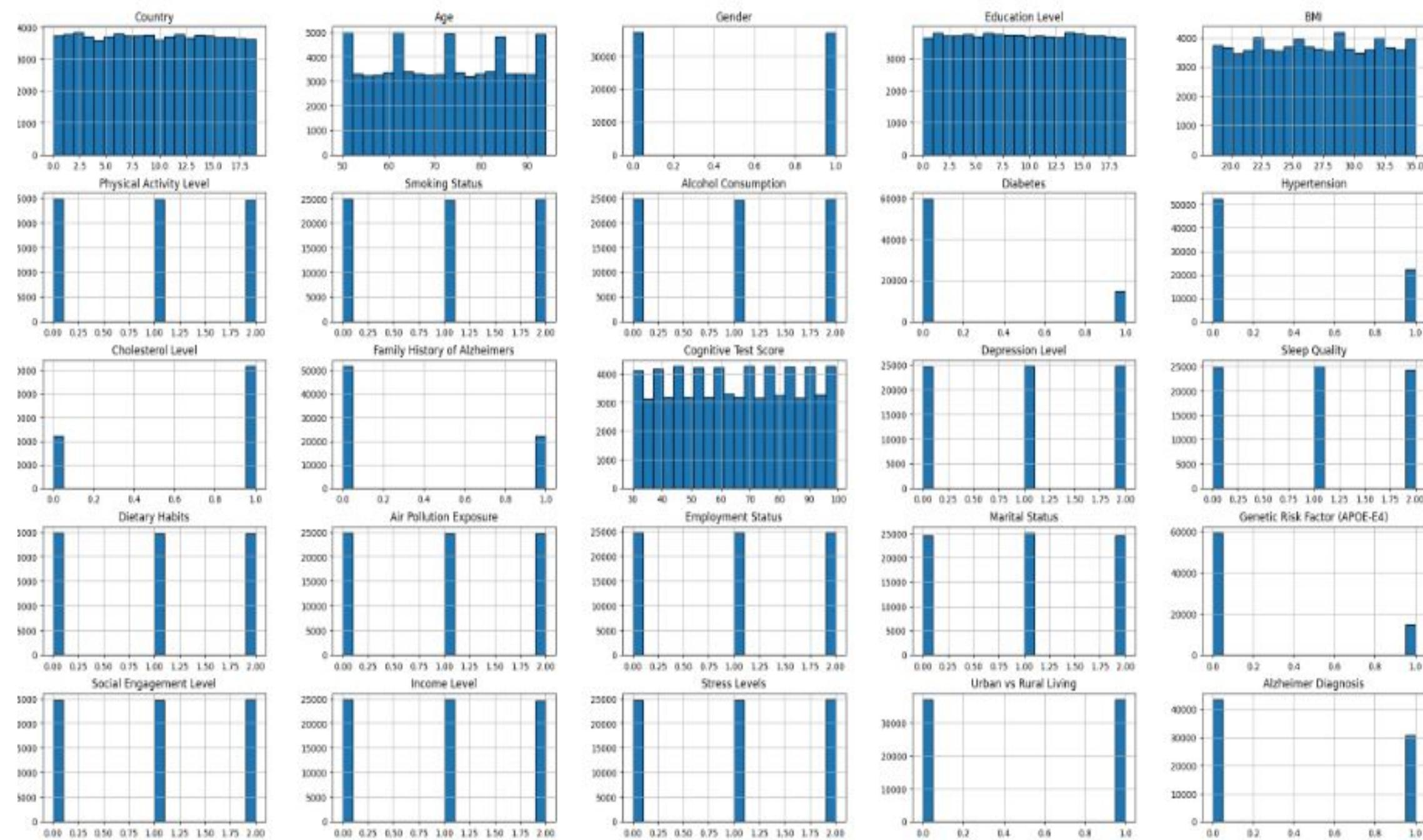
LabelEncoder, 21 features categóricos

```
1 Código → 1 TEXTO
  1 for column in string_columns:
  2     unique_values = data_alz_copy[column].unique()
  3     print(f"Valores únicos transformados en '{column}': {unique_values}")

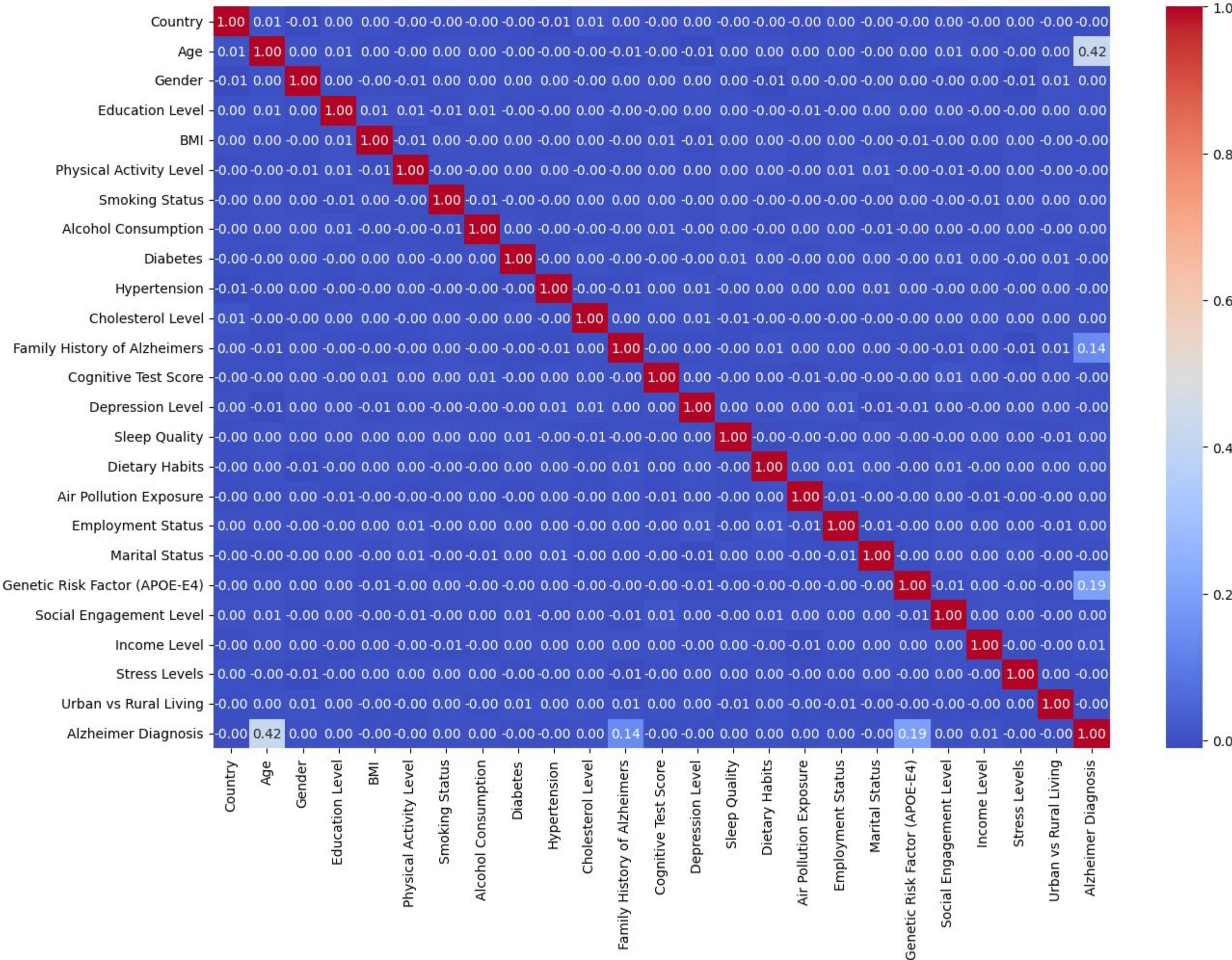
  Valores únicos transformados en 'Country': [12 17 18 19 11 7 1 5 14 16 9 15 0 6 2 4 3 8 13 10]
  Valores únicos transformados en 'Gender': [1 0]
  Valores únicos transformados en 'Physical Activity Level': [1 0 2]
  Valores únicos transformados en 'Smoking Status': [1 0 2]
  Valores únicos transformados en 'Alcohol Consumption': [1 0 2]
  Valores únicos transformados en 'Diabetes': [0 1]
  Valores únicos transformados en 'Hypertension': [0 1]
  Valores únicos transformados en 'Cholesterol Level': [0 1]
  Valores únicos transformados en 'Family History of Alzheimers': [0 1]
  Valores únicos transformados en 'Depression Level': [2 0 1]
  Valores únicos transformados en 'Sleep Quality': [2 0 1]
  Valores únicos transformados en 'Dietary Habits': [2 1 0]
  Valores únicos transformados en 'Air Pollution Exposure': [1 0 2]
  Valores únicos transformados en 'Employment Status': [1 0 2]
  Valores únicos transformados en 'Marital Status': [0 2 1]
  Valores únicos transformados en 'Genetic Risk Factor (APOE-E4)': [0 1]
  Valores únicos transformados en 'Social Engagement Level': [0 2 1]
  Valores únicos transformados en 'Income Level': [0 2 1]
  Valores únicos transformados en 'Stress Levels': [1 0 2]
  Valores únicos transformados en 'Urban vs Rural Living': [1 0]
  Valores únicos transformados en 'Alzheimer Diagnosis': [0 1]
```

## Análisis Univariado

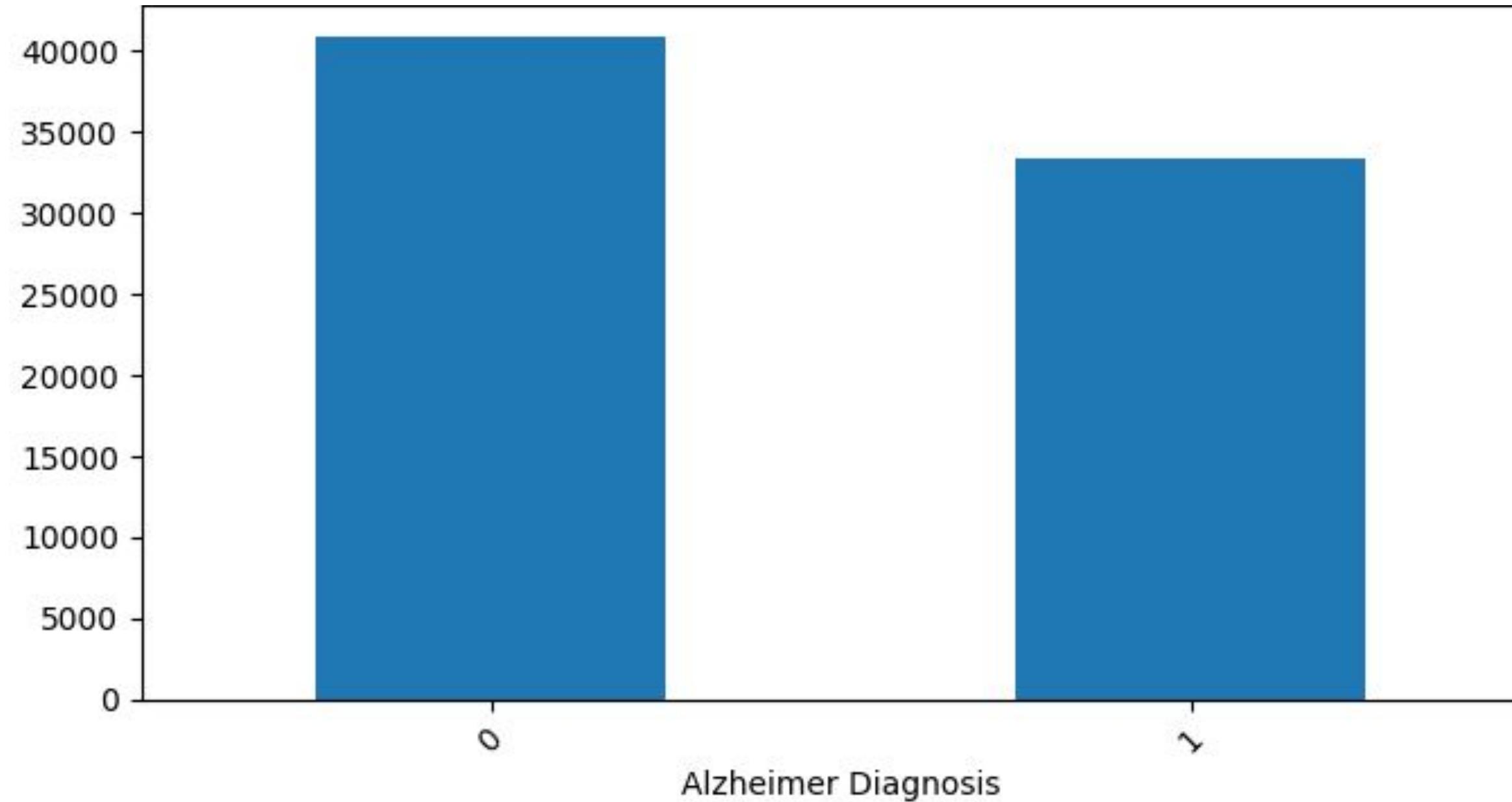
### Histogramas



Matriz de Correlación



# Distribución de la variable a predecir



**Distribución de clases antes del balanceo:**

**Alzheimer Diagnosis**

0	43510
1	30673

Balanceo ?

Proporción de 1:2:No

**Desbalance significativo de clases:** No

verificamos sobremuestreo SMOTE, submuestreo y ADASYN—No cambio el Accuracy

# Escenarios

```
scenarios = [
    "Scenario 1": data_alzheimer.drop(['Country', 'Diabetes', 'Hypertension', 'Cholesterol Level',
                                      'Genetic Risk Factor (APOE-E4)', 'Alzheimer Diagnosis'], axis=1),

    "Scenario 2": data_alzheimer.drop(['Country', 'Education Level', 'Physical Activity Level', 'Smoking Status',
                                      'Alcohol Consumption', 'Dietary Habits', 'Air Pollution Exposure',
                                      'Employment Status', 'Marital Status', 'Social Engagement Level',
                                      'Income Level', 'Stress Levels', 'Urban vs Rural Living', 'Alzheimer Diagnosis'],
                                      axis=1),

    "Scenario 3": data_alzheimer.drop(['Country', 'Age', 'Gender', 'BMI', 'Physical Activity Level', 'Smoking Status',
                                      'Alcohol Consumption', 'Diabetes', 'Hypertension', 'Cholesterol Level',
                                      'Dietary Habits', 'Air Pollution Exposure', 'Employment Status', 'Marital Status',
                                      'Genetic Risk Factor (APOE-E4)', 'Social Engagement Level', 'Income Level',
                                      'Urban vs Rural Living', 'Alzheimer Diagnosis'], axis=1),

    "Scenario 4": data_alzheimer.drop(['Country', 'Age', 'Gender', 'Education Level', 'Diabetes', 'Hypertension',
                                      'Cholesterol Level', 'Cognitive Test Score', 'Depression Level', 'Employment Status',
                                      'Marital Status', 'Genetic Risk Factor (APOE-E4)', 'Income Level', 'Alzheimer Diagnosis'],
                                      axis=1),

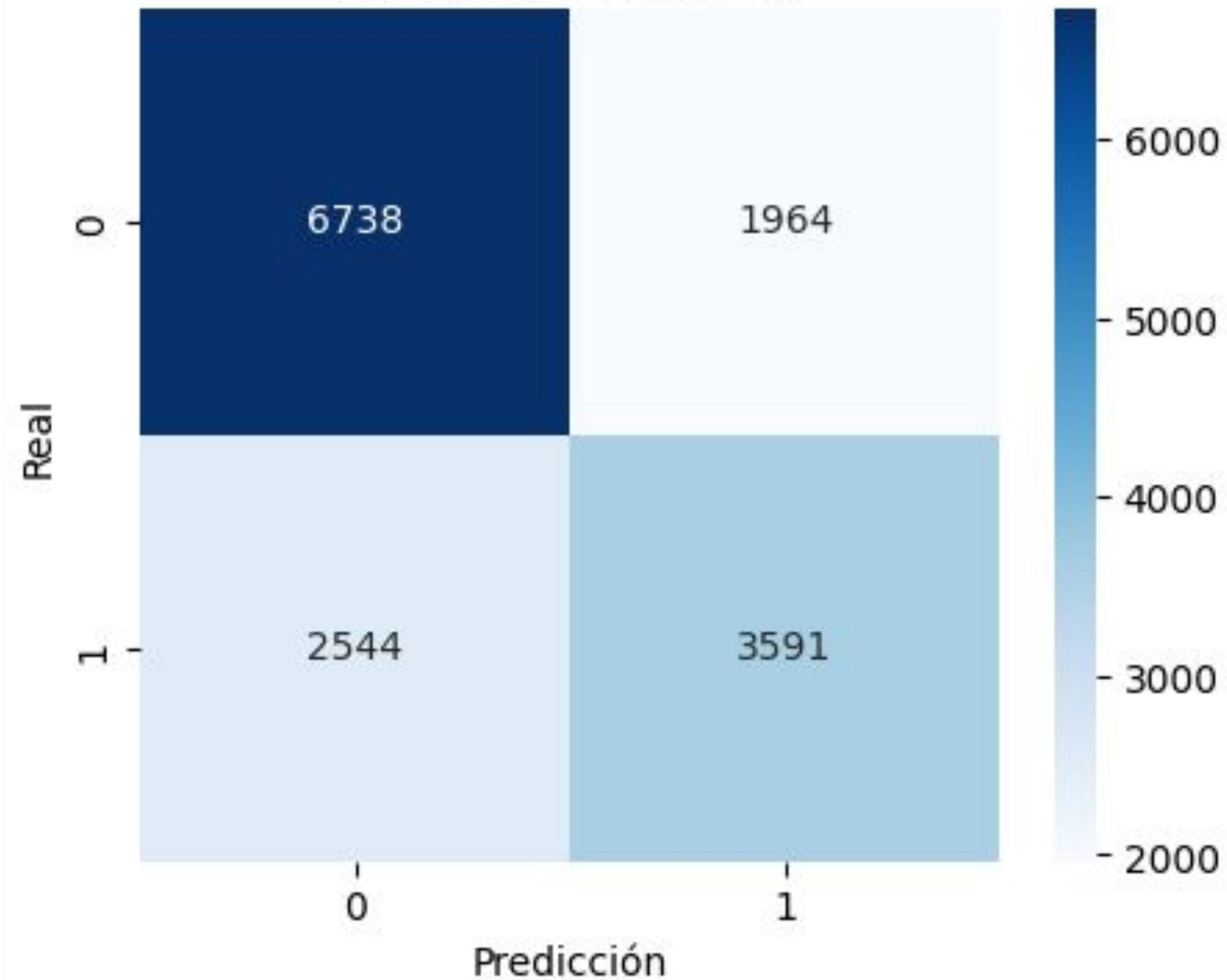
    "Scenario 5": data_alzheimer.drop(['Country', 'BMI', 'Physical Activity Level', 'Smoking Status', 'Alcohol Consumption',
                                      'Diabetes', 'Hypertension', 'Cholesterol Level', 'Cognitive Test Score',
                                      'Depression Level', 'Sleep Quality', 'Dietary Habits', 'Air Pollution Exposure',
                                      'Genetic Risk Factor (APOE-E4)', 'Stress Levels', 'Alzheimer Diagnosis'], axis=1),

    "Scenario 6": data_alzheimer.drop(['Country', 'Alzheimer Diagnosis'], axis=1),

    "Scenario 7": data_alzheimer.drop(['Alzheimer Diagnosis'], axis=1)
```

Algoritmos Utilizados: Logistic Regression, SVM, Decision Tree, Random Forest, XGBoost, Naive Bayes, KNeighbors, GradientBoostingClassifier, LGBMClassifier

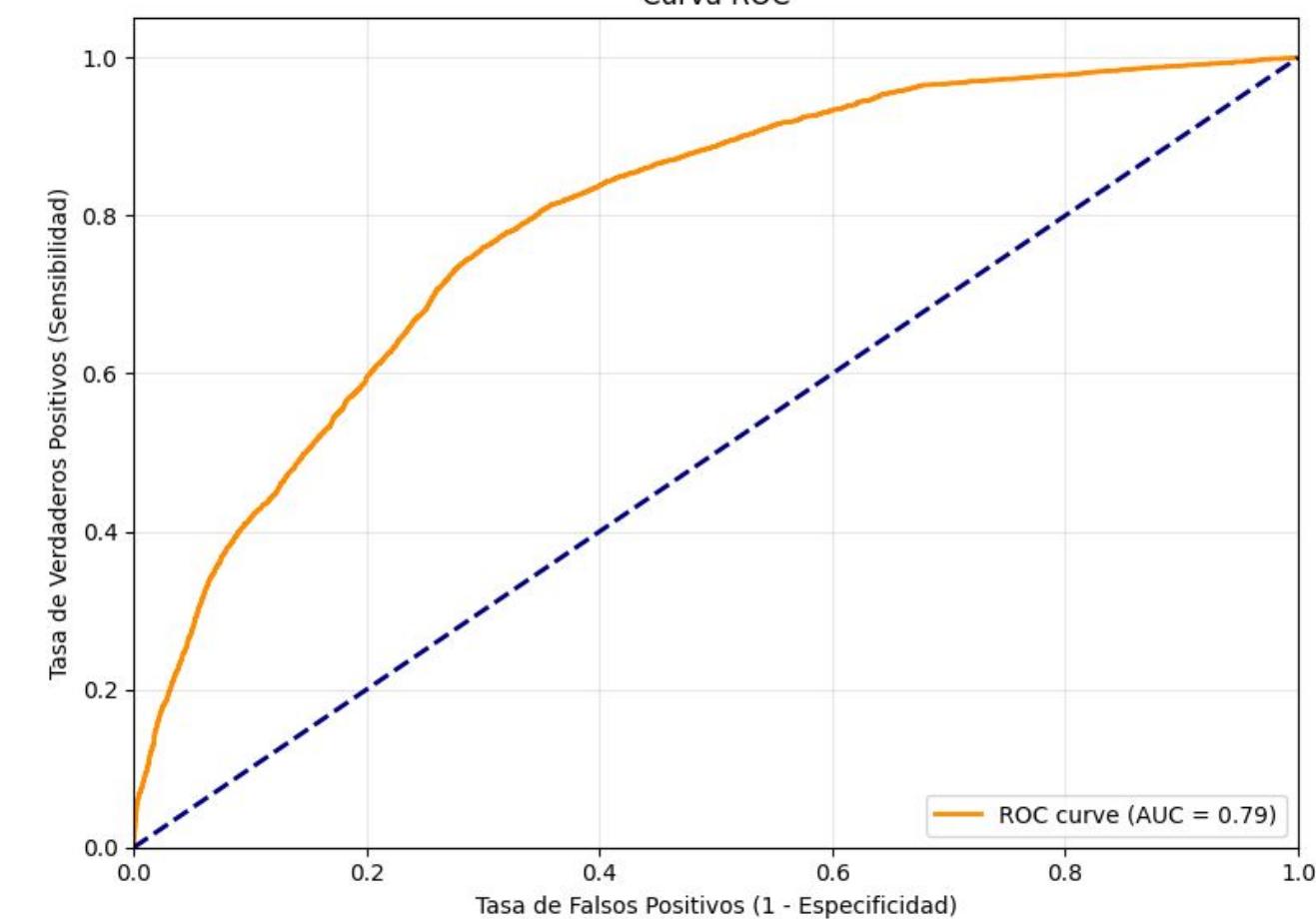
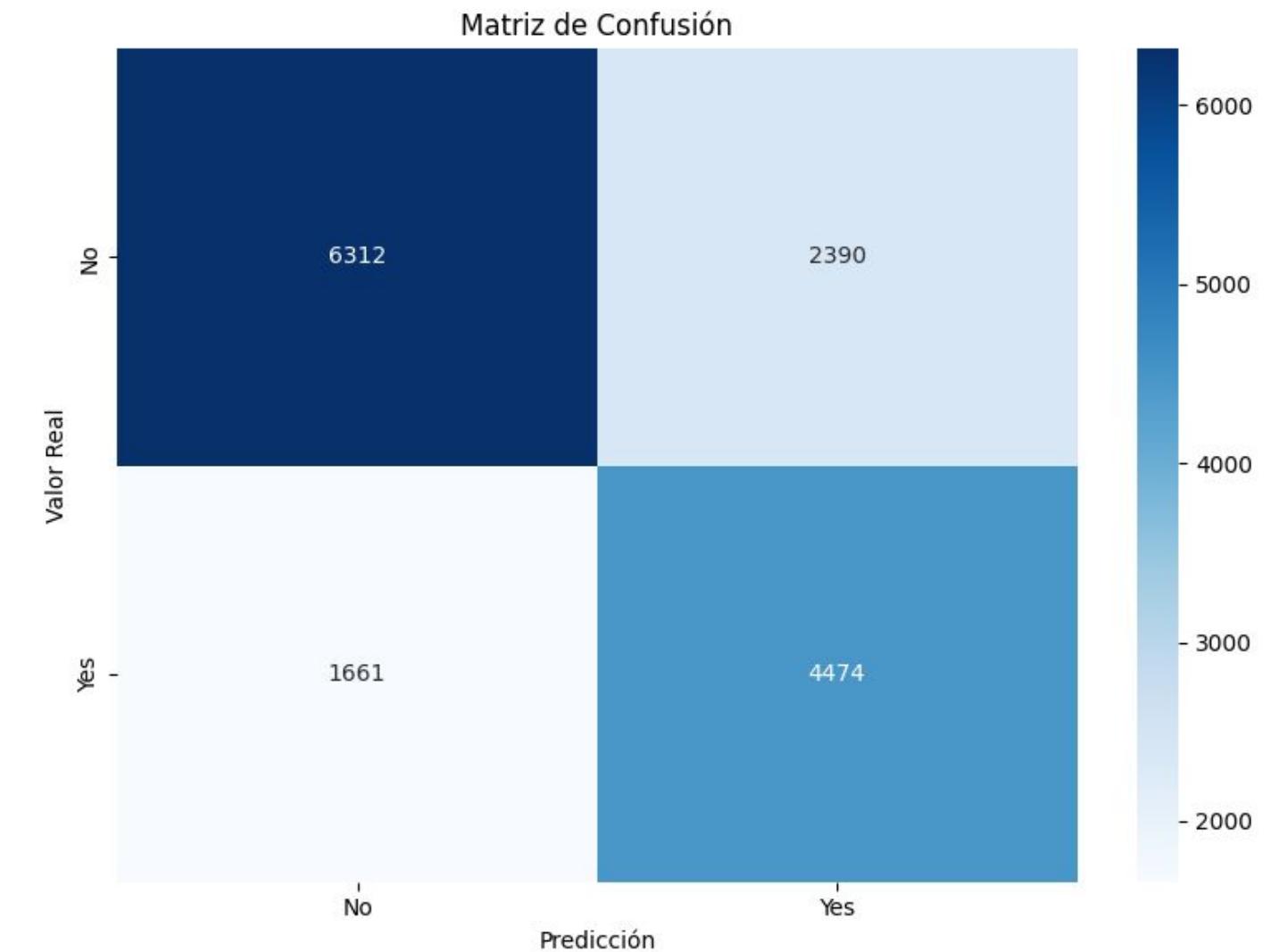
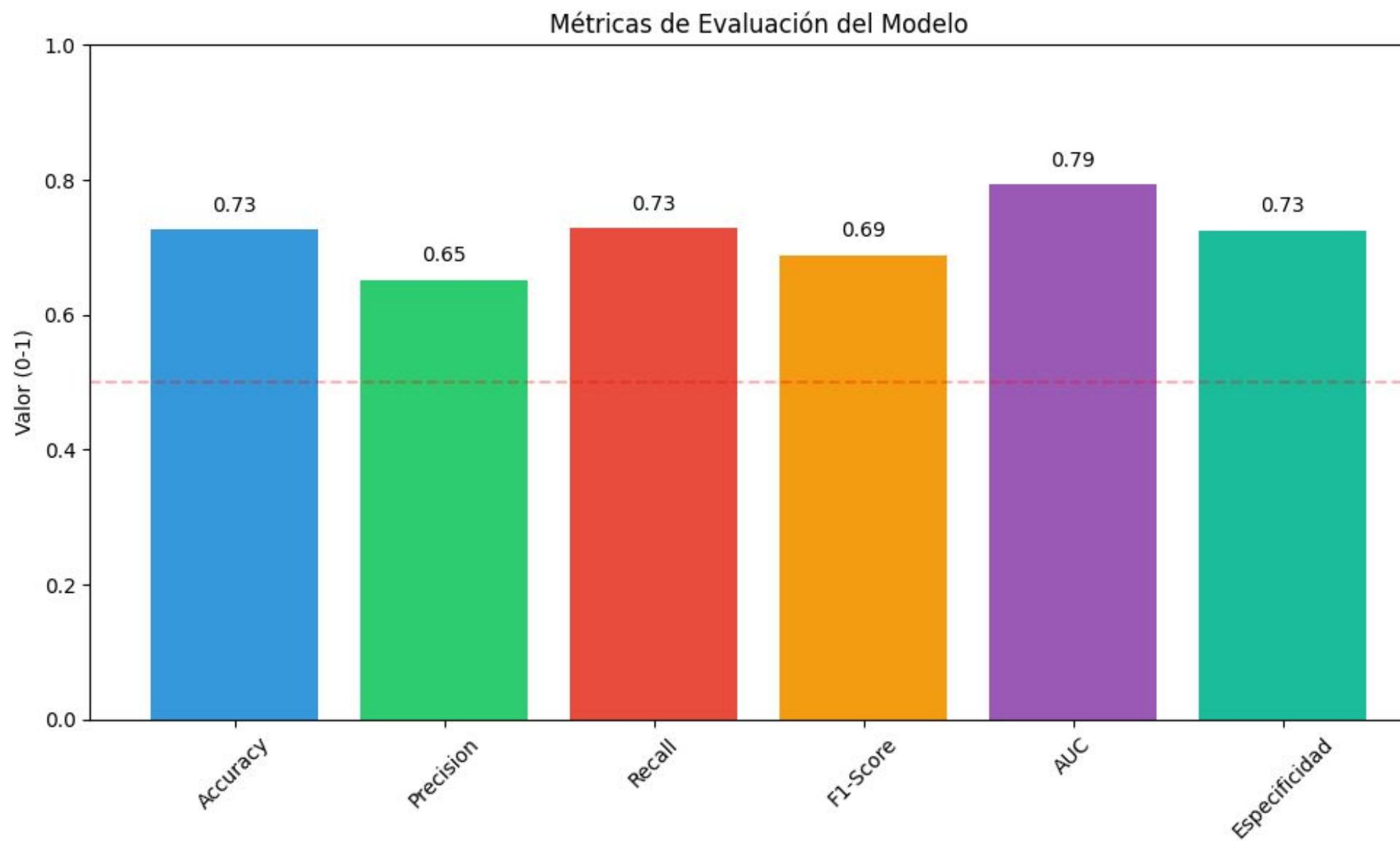
## Matriz de Confusión



# Algoritmos con mejor resultado

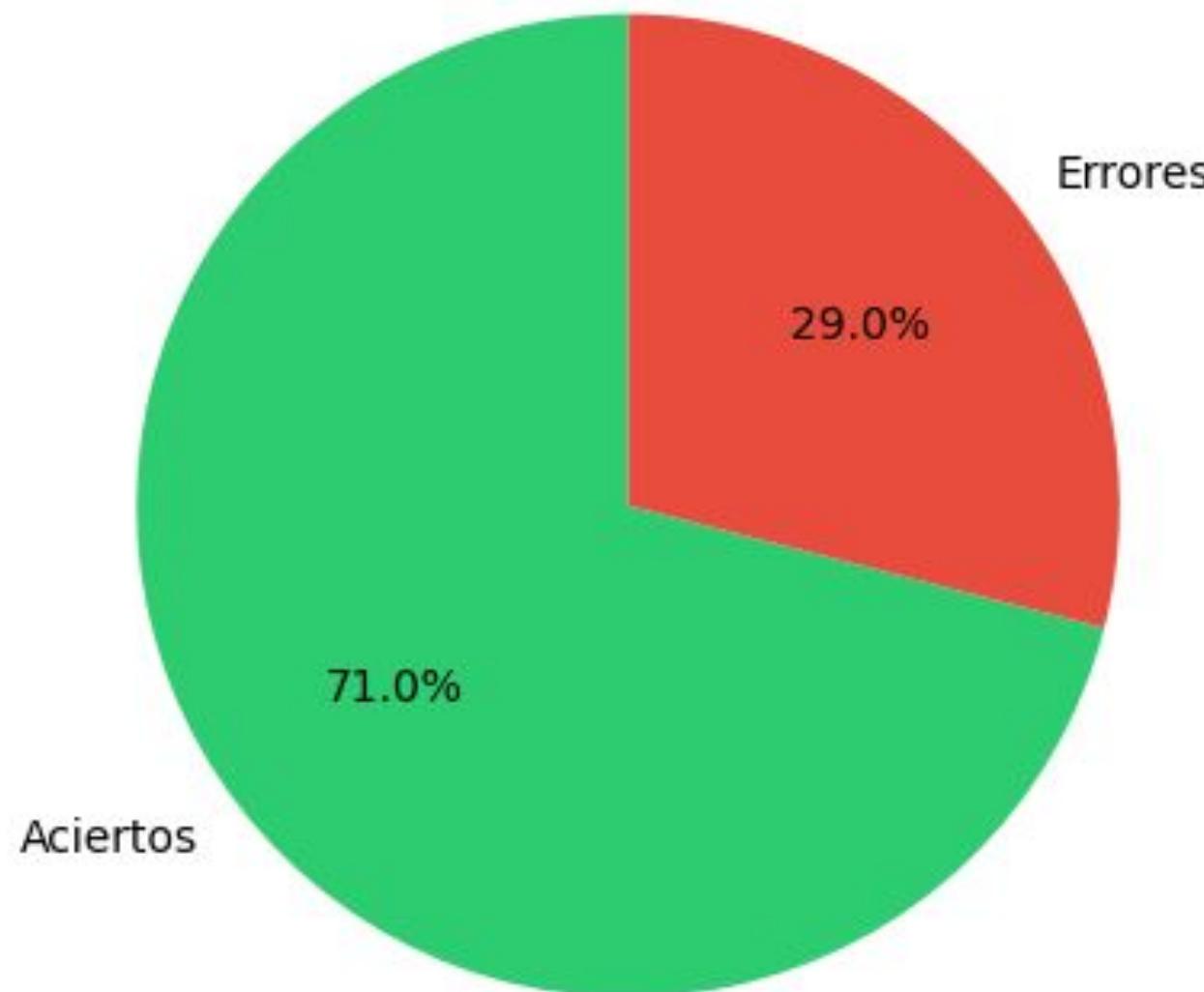
Modelo	Accuracy (Balanceado)	Accuracy (Sin Balanceo)	F1-score (Balanceado)	F1-score (Sin Balanceo)	ROC-AUC (Balanceado)	ROC-AUC (Sin Balanceo)
<b>Random Forest</b>	0.7077	0.7167	0.7097	0.7157	0.7718	0.7880
<b>Gradient Boosting</b>	0.7096	<b>0.7273</b>	0.7118	<b>0.7284</b>	0.7807	<b>0.8031</b>
<b>XGBoost</b>	0.6983	0.7189	0.7004	0.7187	0.7683	0.7908
<b>LightGBM</b>	<b>0.7109</b>	<b>0.7289</b>	<b>0.7130</b>	<b>0.7295</b>	<b>0.7829</b>	<b>0.8016</b>

# Algoritmo elegido : CatBoost

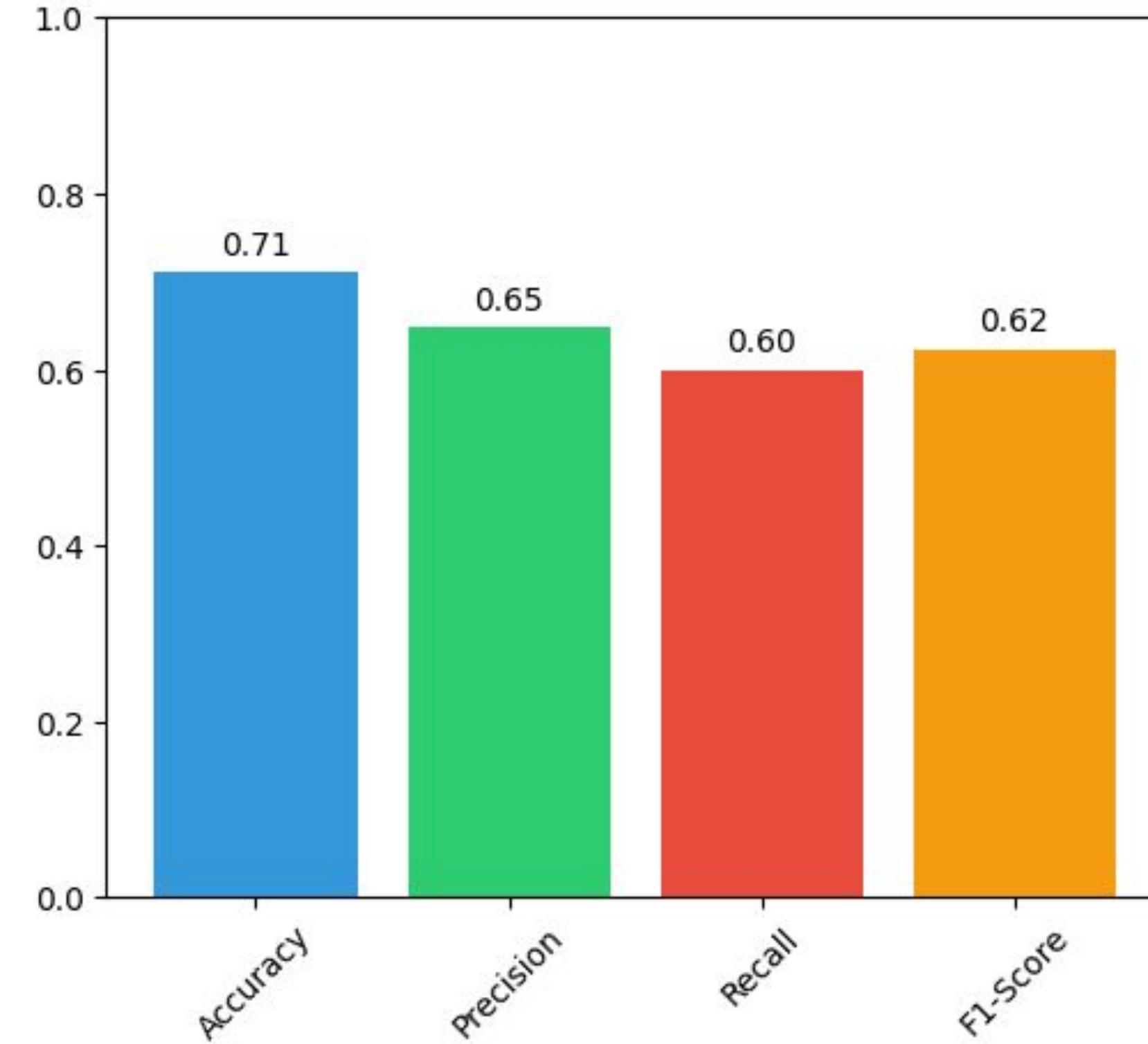


# Evaluación del modelo

Proporción de Aciertos vs Errores



Métricas del Modelo en Datos de Prueba



## Machine learning prediction of incidence of Alzheimer's disease using large-scale administrative health data

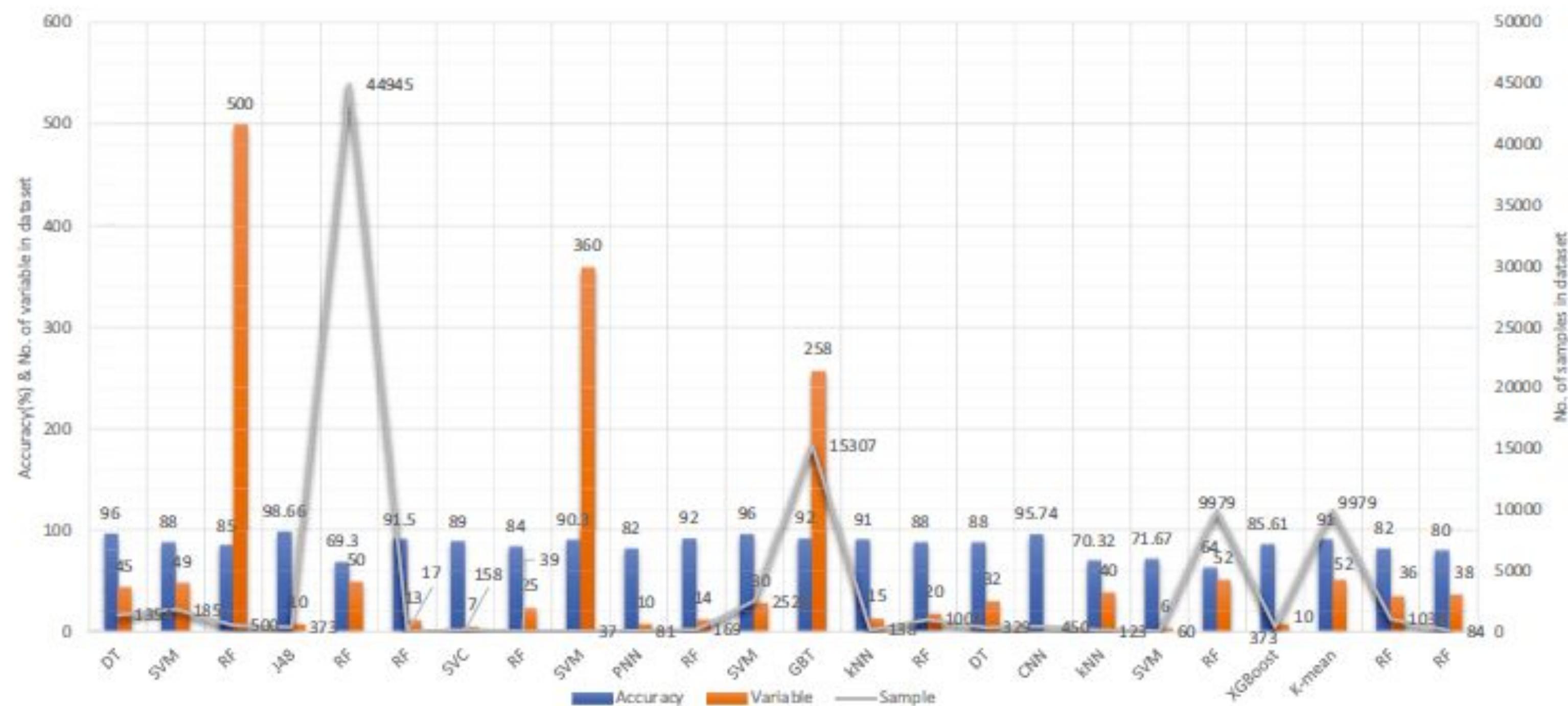
Ji Hwan Park<sup>3,11</sup>, Han Eol Cho<sup>1,2,11</sup>, Jong Hun Kim<sup>3</sup>, Melanie M. Wall<sup>4</sup>, Yaakov Stern<sup>4,5</sup>, Hyunsun Lim<sup>6</sup>, Shinjae Yoo<sup>1</sup>, Hyoung Seop Kim<sup>1,7</sup> and Jiook Cha<sup>1,8,9,10</sup>

**Table 2.** Performance of AD predictive models trained on NHIS-NSC by using balanced samples.

Sample	Subsequent years of incidence predicted <sup>b</sup>	Classifier	Accuracy	AUC	Sensitivity	Specificity
Probable AD (AD/non-AD 2026/2026)	0 year	LR	0.736	0.783	0.689	0.783
		SVM	0.734	0.794	0.652	0.816
		RF	0.788	0.850 <sup>a</sup>	0.723	0.853
	1 year	LR	0.663	0.697	0.634	0.692
		SVM	0.661	0.691	0.592	0.729
		RF	0.688	0.759 <sup>a</sup>	0.609	0.767
	2 year	LR	0.643	0.672	0.633	0.654
		SVM	0.645	0.68	0.58	0.709
		RF	0.638	0.693 <sup>a</sup>	0.564	0.713
	3 year	LR	0.61	0.635	0.557	0.663
		SVM	0.597	0.644 <sup>a</sup>	0.427	0.767
		RF	0.581	0.609	0.505	0.657
	4 year	LR	0.611	0.644	0.516	0.707
		SVM	0.601	0.641	0.465	0.738
		RF	0.641	0.683 <sup>a</sup>	0.603	0.679

# Machine Learning for Dementia Prediction: A Systematic Review and Future Research Directions

Ashir Javeed<sup>1,2</sup> · Ana Luiza Dallora<sup>2</sup> · Johan Sanmartin Berglund<sup>2</sup> · Arif Ali<sup>3</sup> · Liaqata Ali<sup>4</sup> · Peter Anderberg<sup>2,5</sup>



**Fig. 6** Accuracy comparison of different ML models based on clinical-variable modality

# Implementación y Aplicabilidad

Desarrollo modelo

PKL

Desarrollo web

Uso

01



Clinicas y  
hospitales

02



Consulta  
externa

03



WEB

04



Publicaciones  
y aprendizaje

→ <https://github.com/AndresPatinol/alzheimers-prediction>

→ <https://alzheimer-front.onrender.com/>

**NeuroCheck AI**

**Inicio Diagnosticar**

## Descubre NeuroCheck AI

Un modelo de inteligencia artificial diseñado para detectar signos tempranos de Alzheimer y proporcionar información útil sobre tu salud cognitiva.

[Comenzar Diagnóstico](#)



### ¿Cómo Funciona?

1. Sube tu Información

2. Análisis con IA

3. Obtén tu Resultado

### Inicia tu Diagnóstico

Completa el formulario para que nuestro modelo de inteligencia artificial pueda analizar tu salud cognitiva y detectar signos tempranos de Alzheimer.



Completa la información

Nombre

Escribe tu nombre

Edad

Escribe tu edad

Activar Windows

Ve a Configuración para activar Windows.

### Prevención y Cuidado

Aunque el Alzheimer no tiene cura, estudios han demostrado que ciertos hábitos pueden reducir el riesgo. Aquí te dejamos algunos consejos:

- 🧠 Mantén tu mente activa con ejercicios cognitivos.
- 💡 Lleva una dieta saludable, rica en antioxidantes.
- 🏃 Realiza actividad física regularmente.
- 😴 Duerme lo suficiente para mejorar la memoria.
- 🤝 Mantén relaciones sociales activas.

# Conclusiones

- El desarrollo de modelos predictivos para enfermedades como el Alzheimer es un campo desafiante. Un accuracy de 0.73 puede ser un punto de partida, pero es fundamental considerar otras métricas para evaluar la utilidad real del modelo, especialmente en el contexto de la detección temprana.
- Para nuestro objetivo es muy importante el Recall con el fin de detectar la mayor cantidad de pacientes enfermos posibles que puedan ser evaluados para una confirmación del diagnóstico.
- Requerimos ampliar datos y uso de hiperparámetros para mejorar las métricas





# Gracias

