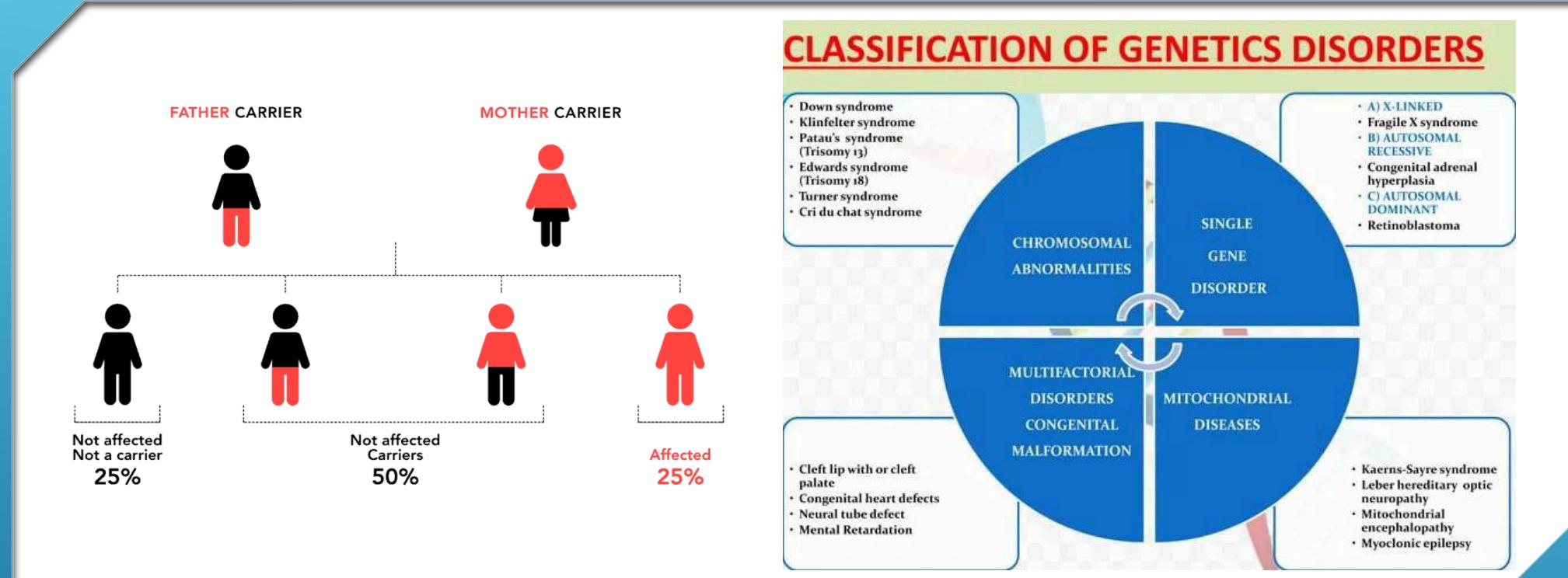


# PREDICCIÓN DE DESÓRDENES GENÉTICOS EN NIÑOS: UN ENFOQUE DE MACHINE LEARNING

M. En C. Israel Solano

Proyecto Final: Diplomado en Ciencia de Datos con Python 2024



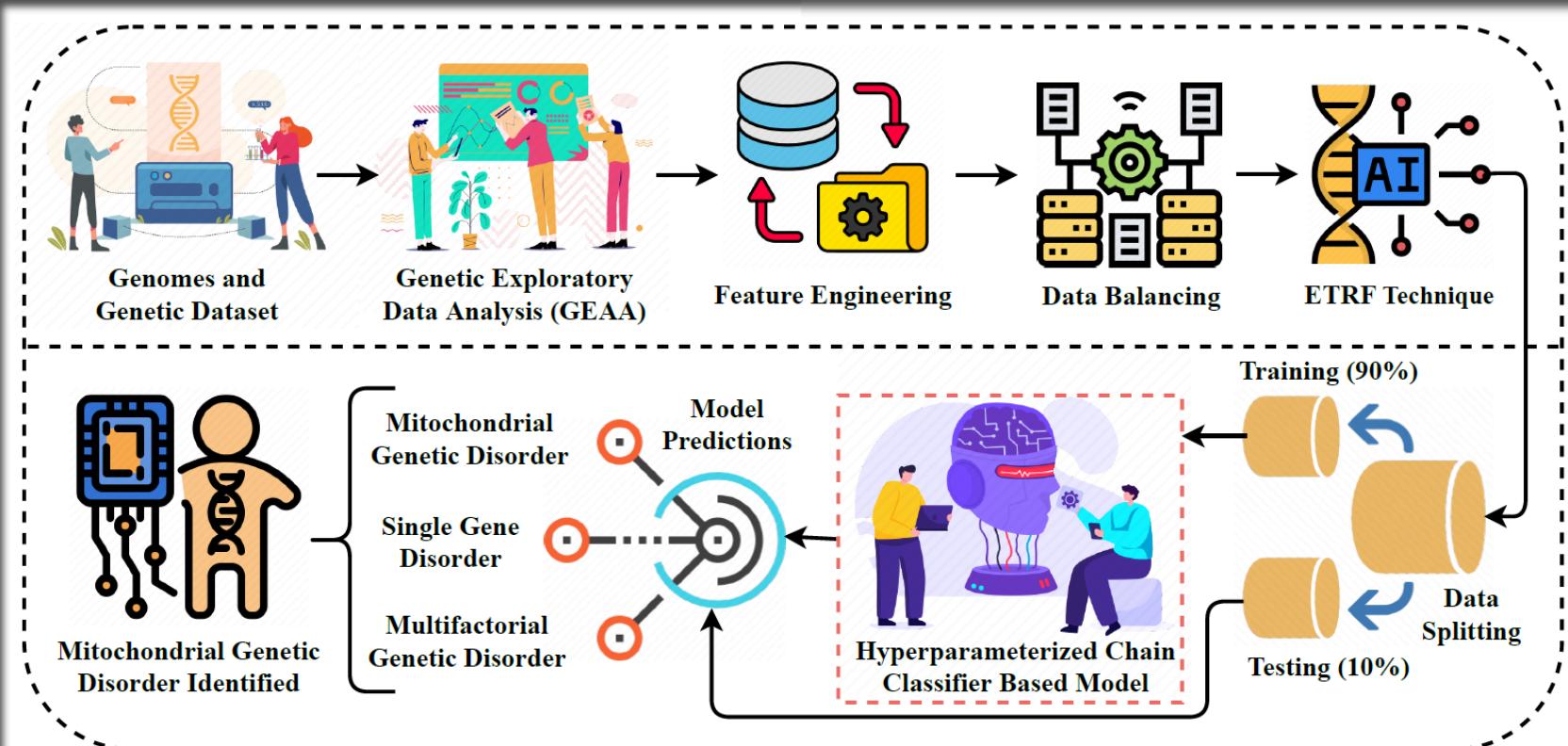
# TIPOS DE DESÓRDENES GENÉTICOS

# DEFINICIÓN DEL PROBLEMA

- ▶ Con un conjunto de datos médicos suficientemente grande y de alta calidad, es posible desarrollar un modelo de aprendizaje automático que pueda predecir con precisión los desórdenes genéticos y sus subclases en niños.

# OBJETIVOS

- ▶ Predecir desórdenes genéticos y sus subclases en niños para ayudar en el diagnóstico y el tratamiento temprano de estas enfermedades.
- ▶ Desarrollar un modelo de aprendizaje automático que pueda hacer estas predicciones con un alto grado de precisión.



Article

## Predicting Genetic Disorder and Types of Disorder Using Chain Classifier Approach

Ali Raza <sup>1</sup>, Furqan Rustam <sup>2</sup>, Hafeez Ur Rehman Siddiqui <sup>1</sup>, Isabel de la Torre Diez <sup>3,\*</sup>,  
Begoña García-Zapirain <sup>4</sup>, Ernesto Lee <sup>5</sup> and Imran Ashraf <sup>6,\*</sup>

Reference	Year	Technique	Training Time (s)	Macro Accuracy (%)	Hamming Loss	$\alpha$ -Evaluation Score (%)
[63]	2020	SVM	7.10	73	0.22	88
[64]	2020	KNN	0.01	70	0.25	86
[65]	2020	KNN	0.01	70	0.25	86
[66]	2020	RF	2.48	82	0.14	90
[67]	2021	KNN	0.01	70	0.25	86
Proposed	2022	ETRF + XGB	3.59	84	0.12	92

# METODOLOGÍA GENERAL

# DESCRIPCIÓN DEL CONJUNTO DE DATOS

- ▶ El conjunto de datos contiene información médica de niños con desordenes genéticos y sus padres

▶ **Origen del conjunto de datos:** Proviene de un desafío de Machine Learning, centrado en trastornos genéticos en niños.

▶ **Tamaño y diversidad:** 22,000 registros únicos con una amplia gama de características médicas y genéticas.

▶ Este tipo de datos es multiclase y multietiqueta, lo que conlleva desafíos para manejarlos de manera efectiva.

▶ Permite trabajar en un problema con implicaciones reales en medicina genética.

Column name	Column description
Patient Id	Represents the unique identification number of a patient
Patient Age	Represents the age of a patient
Genes in mother's side	Represents a gene defect in a patient's mother
Inherited from father	Represents a gene defect in a patient's father
Maternal gene	Represents a gene defect in the patient's maternal side of the family
Paternal gene	Represents a gene defect in a patient's paternal side of the family
Blood cell count (mcL)	Represents the blood cell count of a patient
Patient First Name	Represents a patient's first name
Family Name	Represents a patient's family name or surname
Father's name	Represents a patient's father's name
Mother's age	Represents a patient's mother's name
Father's age	Represents a patient's father's age
Institute Name	Represents the medical institute where a patient was born
Location of Institute	Represents the location of the medical institute
Status	Represents whether a patient is deceased
Respiratory Rate (breaths/min)	Represents a patient's respiratory breathing rate

Heart Rate (rates/min)	Represents a patient's heart rate
Test 1 - Test 5	Represents different (masked) tests that were conducted on a patient
Parental consent	Represents whether a patient's parents approved the treatment plan
Follow-up	Represents a patient's level of risk (how intense their condition is)
Gender	Represents a patient's gender
Birth asphyxia	Represents whether a patient suffered from birth asphyxia
Autopsy shows birth defect (if applicable)	Represents whether a patient's autopsy showed any birth defects
Place of birth	Represents whether a patient was born in a medical institute or home
Folic acid details (peri-conceptional)	Represents the periconceptional folic acid supplementation details of a patient
H/O serious maternal illness	Represents an unexpected outcome of labor and delivery that resulted in significant short or long-term consequences to a patient's mother
H/O radiation exposure (x-ray)	Represents whether a patient has any radiation exposure history
H/O substance abuse	Represents whether a parent has a history of drug addiction
Assisted conception IVF/ART	Represents the type of treatment used for infertility

History of anomalies in previous pregnancies	Represents whether the mother had any anomalies in her previous pregnancies
No. of previous abortion	Represents the number of abortions that a mother had
Birth defects	Represents whether a patient has birth defects
White Blood cell count (thousand per microliter)	Represents a patient's white blood cell count
Blood test result	Represents a patient's blood test results
Symptom 1 - Symptom 5	Represents (masked) different types of symptoms that a patient had
Genetic Disorder	Represents the genetic disorder that a patient has
Disorder Subclass	Represents the subclass of the disorder

# ANALISIS EXPLORATORIO DE DATOS GENÉTICOS

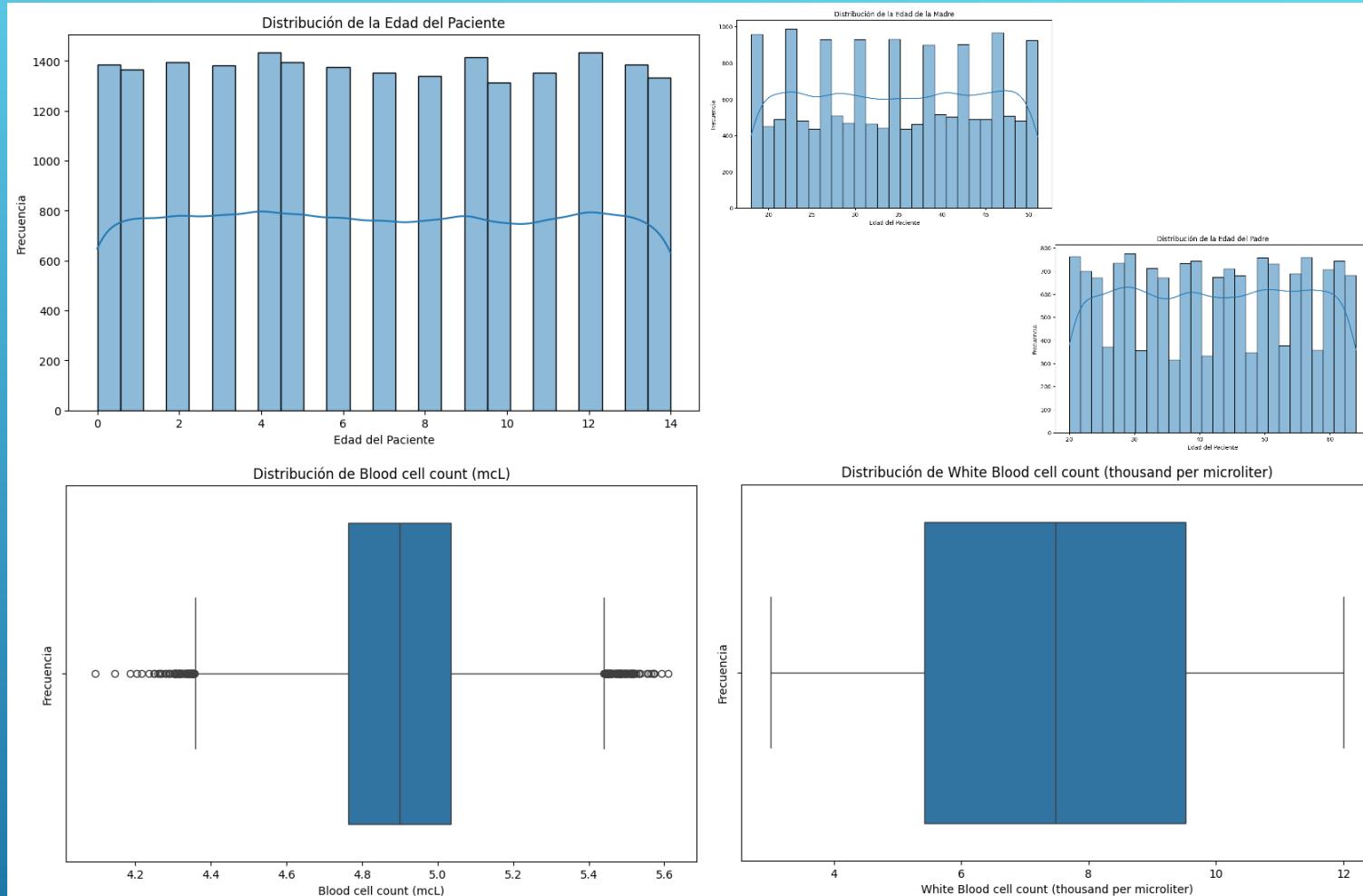
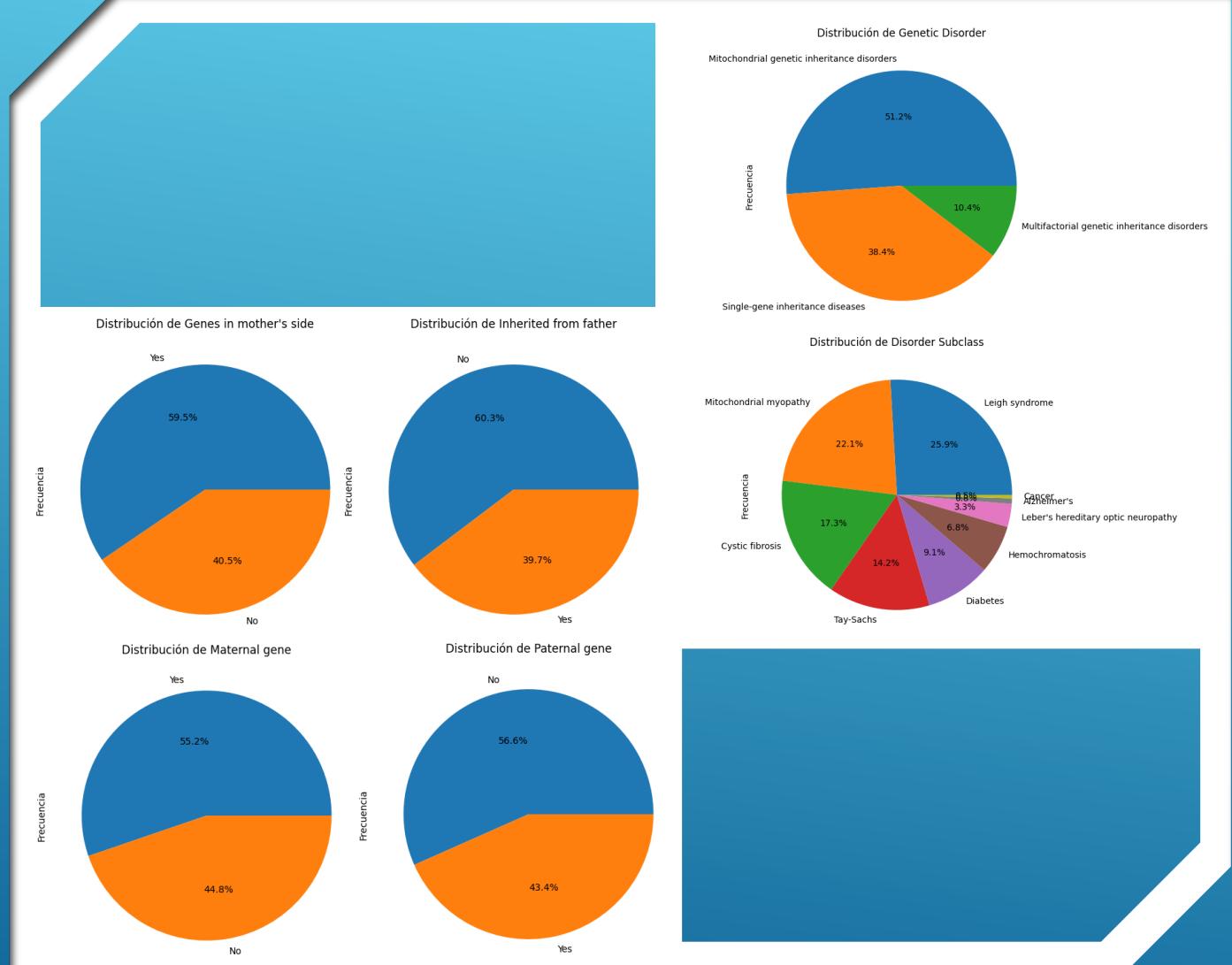
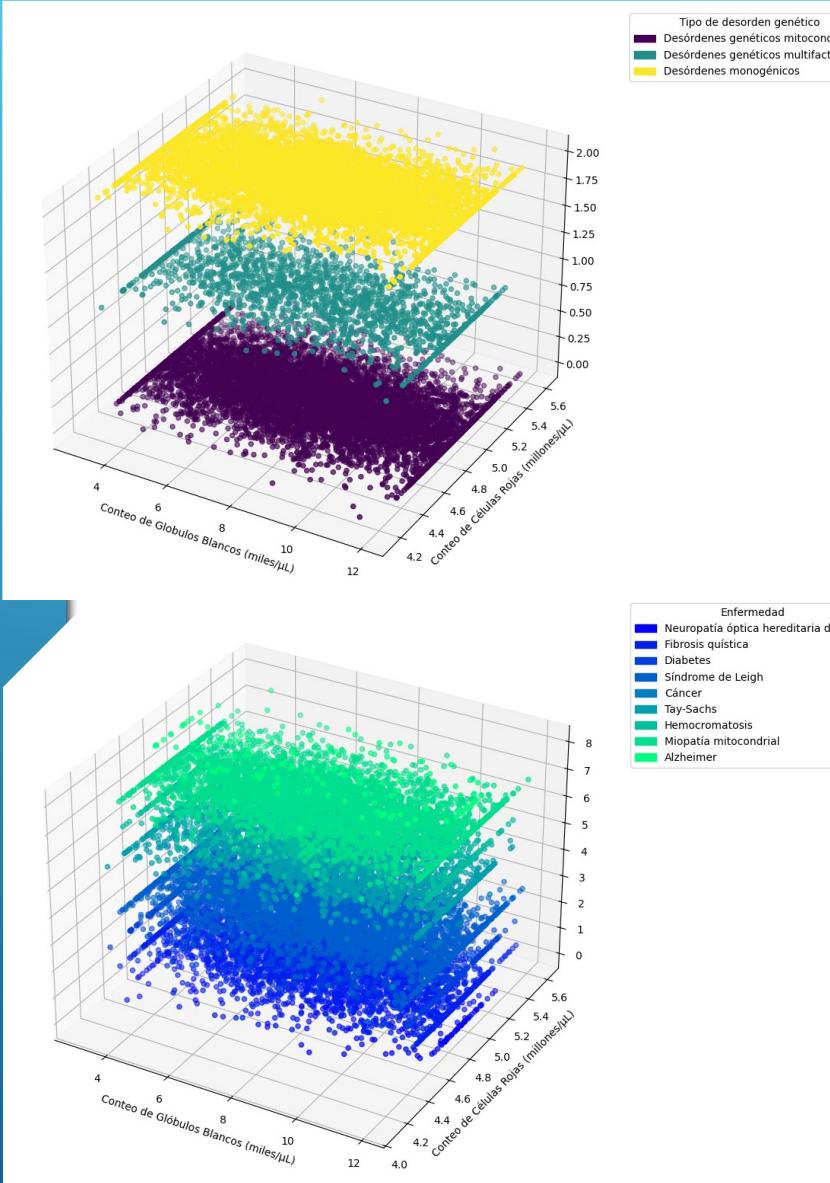


Tabla 1. Resumen Estadístico de Variables Numéricas

	Edad del Paciente	Edad de la madre	Edad del padre	Numero de abortos previos	Conteo de globulos rojos(mcL)	Conteo de globulos blancos (K/ $\mu$ L)
Registros	20656	16047	16097	19921	22083	19935
Promedio	6.97	34.53	41.97	2.00	4.90	7.49
Desviacion estandar	4.32	9.85	13.04	1.41	0.20	2.65
Min	0.00	18.00	20.00	0.00	4.09	3.00
Max	14.00	51.00	64.00	4.00	5.61	12.00

# ANALISIS EXPLORATORIO DE DATOS GENÉTICOS





#### ► Conteo de glóbulos blancos y trastornos genéticos:

- Cuando el conteo de glóbulos blancos es menor a cero, se encuentran trastornos genéticos de todo tipo. Entre 0 y 2, no hay posibilidad de trastorno mitocondrial (tipo 0), pero sí de trastornos multifactoriales (tipo 1) y de un solo gen (tipo 2).

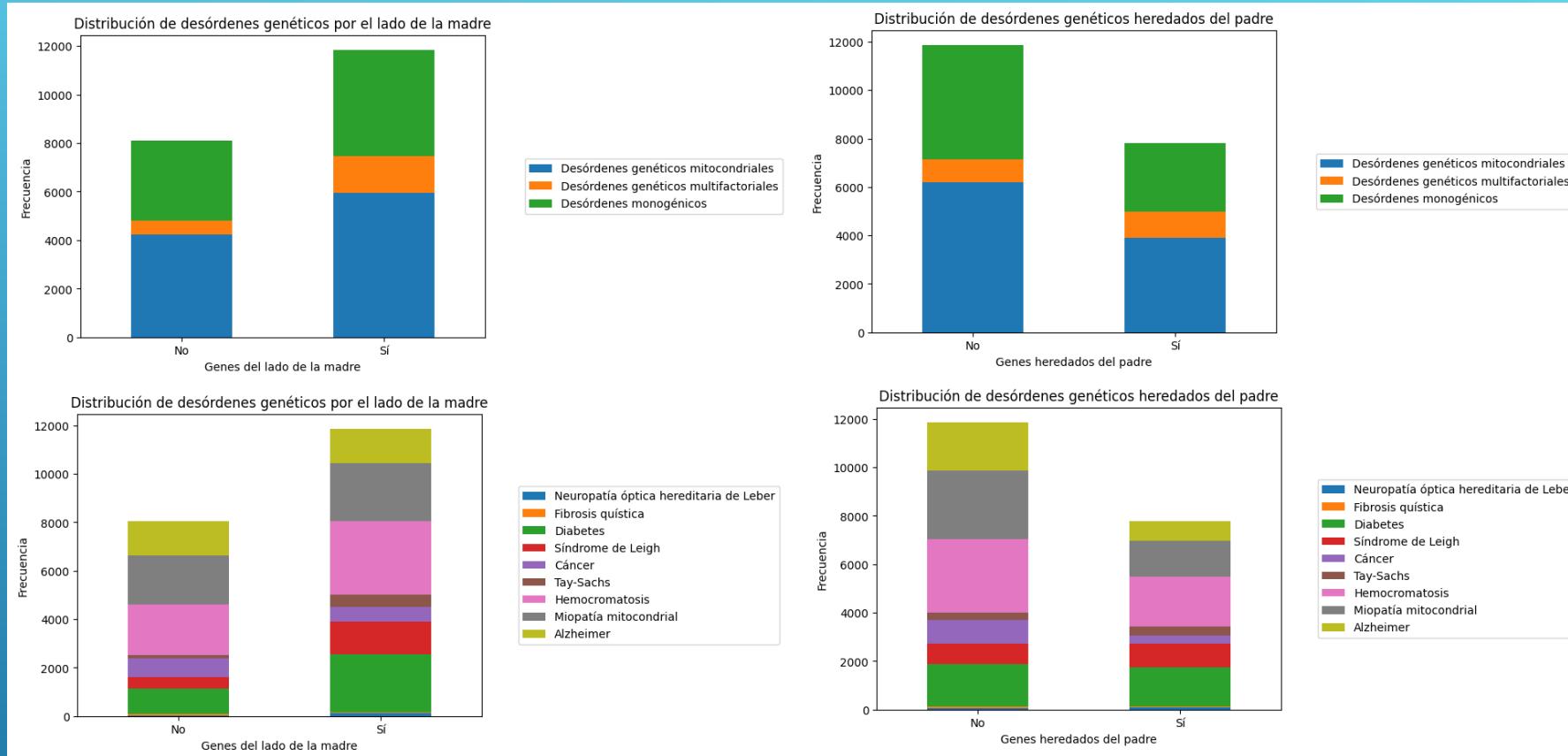
#### ► Conteo de células sanguíneas y trastornos genéticos:

- Valores de conteo de células sanguíneas de 4.2 mCL o menos indican ausencia de trastornos genéticos. Entre 4.3 y 5.6 mCL, se encuentran trastornos genéticos de todo tipo.

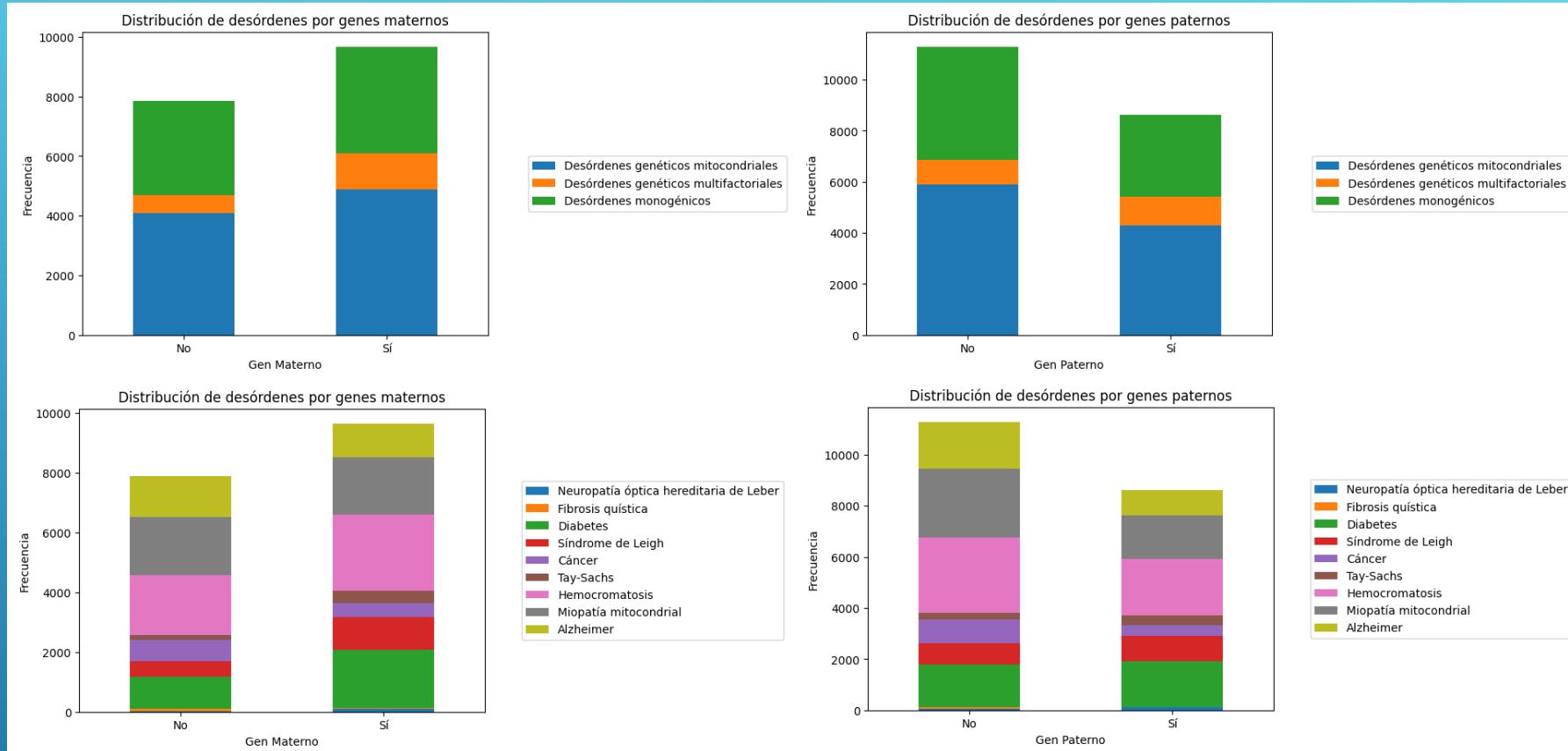
#### ► Subclases de trastornos genéticos:

- No se encuentran subclases de enfermedades cuando el conteo de células sanguíneas está entre 4.2 y 4.4 mCL. La neuropatía óptica hereditaria de Leber se encuentra entre 4.4 y 4.8 mCL. Valores superiores a 4.8 mCL indican la presencia de todas las subclases de trastornos.

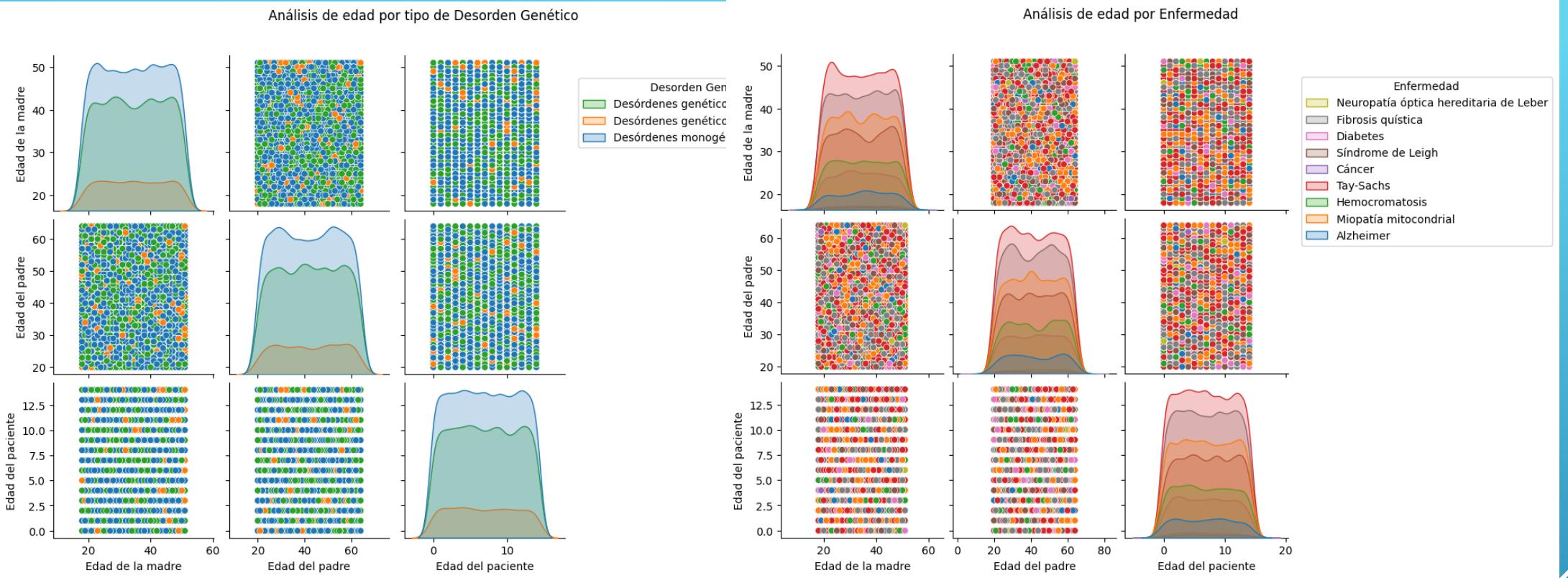
# ANÁLISIS EXPLORATORIO DE DATOS GENÉTICOS



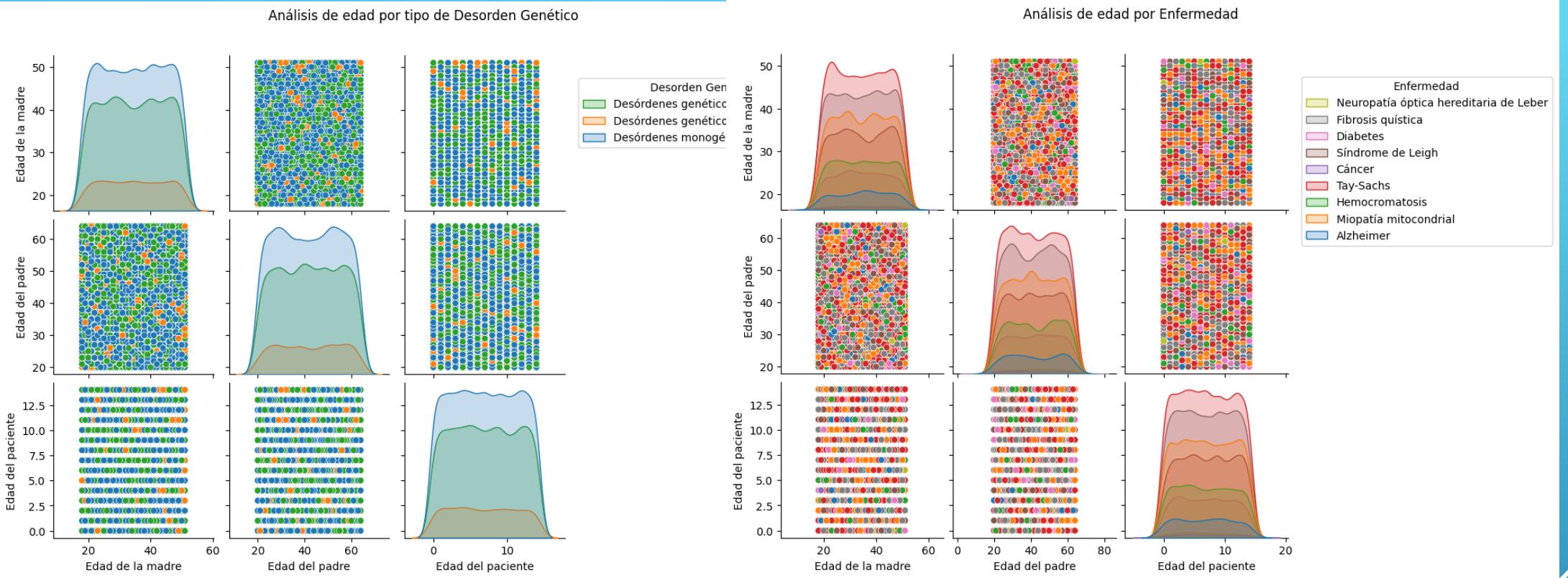
# ANÁLISIS EXPLORATORIO DE DATOS GENÉTICOS



# ANÁLISIS EXPLORATORIO DE DATOS GENÉTICOS

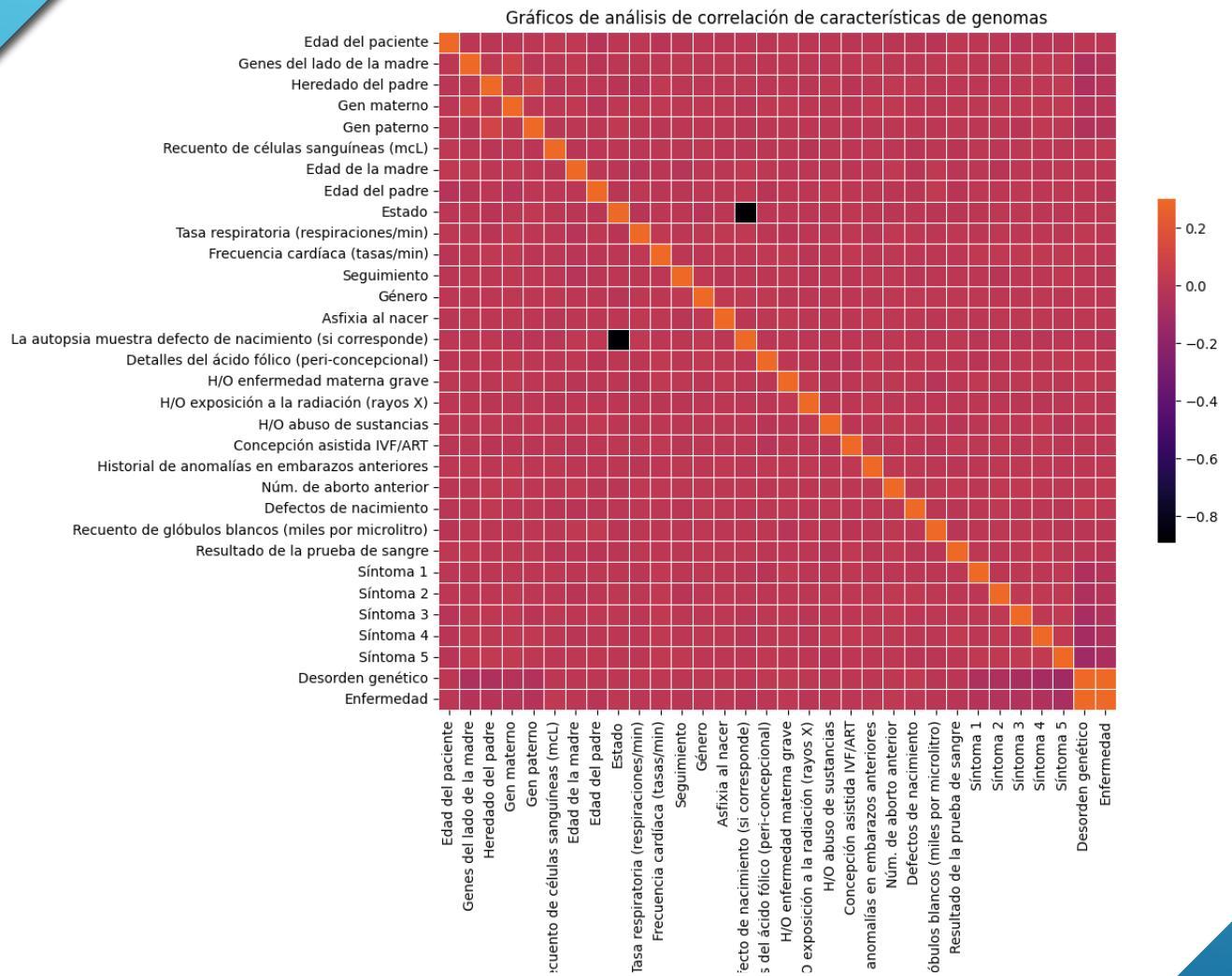


# ANÁLISIS EXPLORATORIO DE DATOS GENÉTICOS

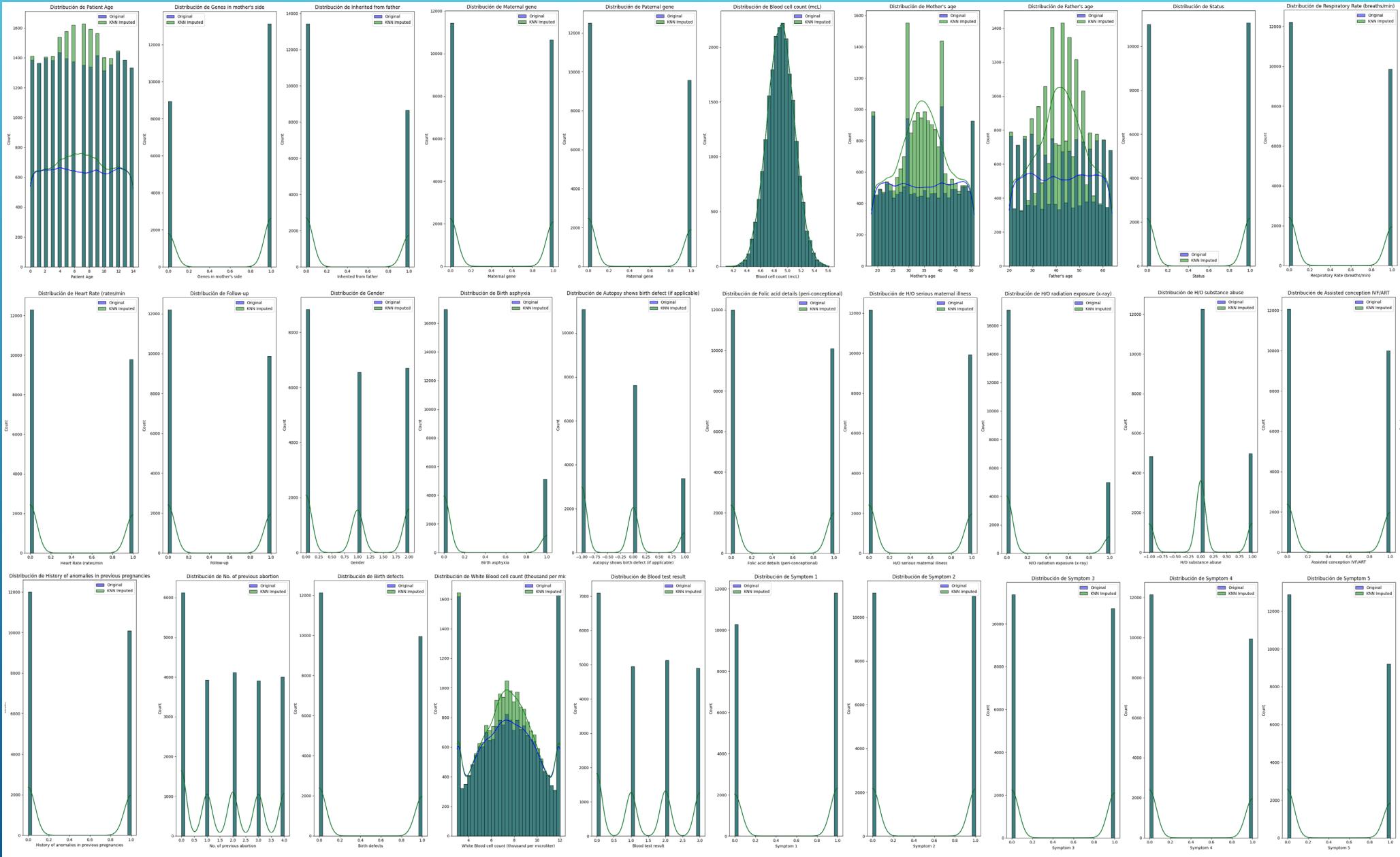


# ANÁLISIS EXPLORATORIO DE DATOS GENÉTICOS

# ANÁLISIS EXPLORATORIO DE DATOS GENÉTICOS

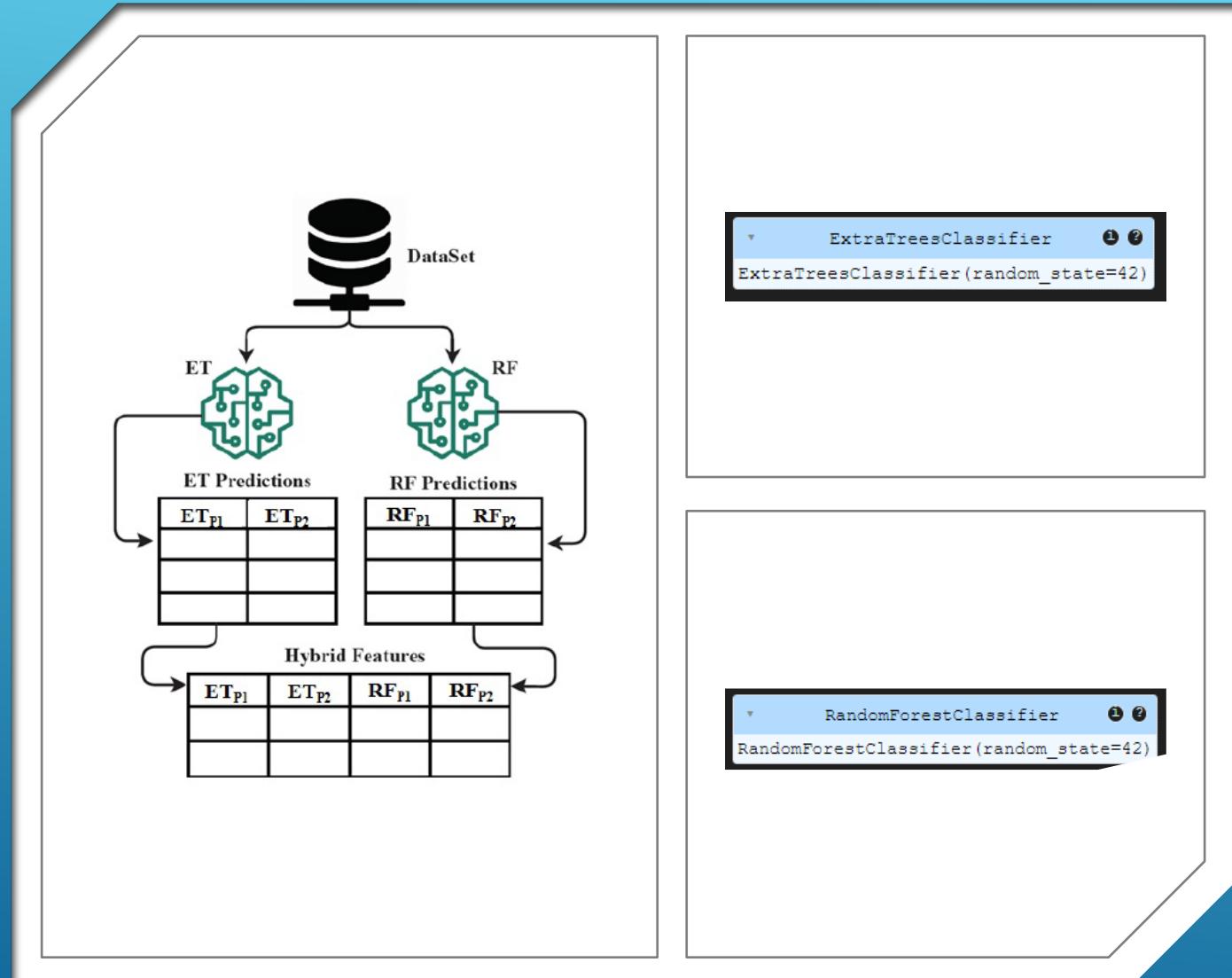


# IMPUTACIÓN DE VALORES FALTANTES



# CONSTRUCCIÓN DEL MODELO ETRF: MODELOS INDEPENDIENTES.

Los resultados mostraron que ninguno de los dos modelos pudo predecir correctamente las etiquetas de las clases.



ET Accuracy (Genetic Disorder): 0.0  
ET Hamming Loss (Genetic Disorder): 1.0  
ET F1 Score (Genetic Disorder): 0.0

ET Accuracy (Disorder Subclass): 0.0  
ET Hamming Loss (Disorder Subclass): 1.0  
ET F1 Score (Disorder Subclass): 0.0

Esta aproximación, no fue adecuada para este problema de clasificación particular.

# CONSTRUCCIÓN DEL MODELO ETRF: MODELOS ANIDADOS

Se implementaron dos modelos de clasificación MultiOutput, utilizando Extra Trees y Random Forest como modelos base.

Estos modelos se entrenaron y luego se evaluaron.

Además, se realizó una búsqueda en cuadrícula manual para optimizar estos modelos.

```
MultiOutputClassifier
MultiOutputClassifier(estimator=ExtraTreesClassifier(random_state=42))
|   estimator: ExtraTreesClassifier
|   ExtraTreesClassifier(random_state=42)
|   |
|   * ExtraTreesClassifier
|       ExtraTreesClassifier(random_state=42)
```

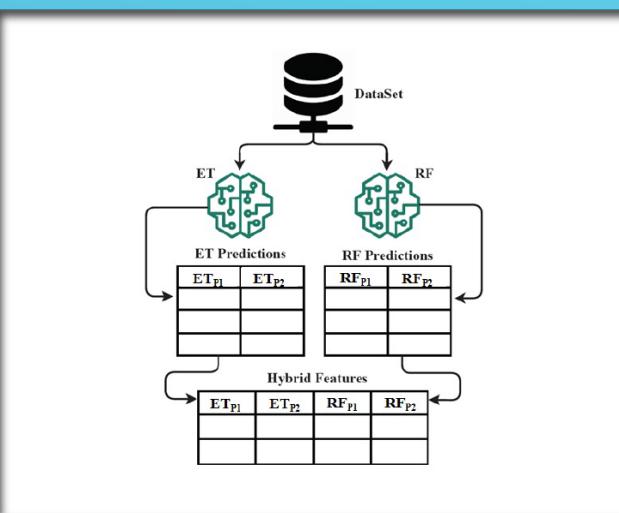


Tabla 2. Construcción de Modelos de Clasificación MultiOutput ETRF					
Modelo	Parámetros	Exactitud Promedio	F1-Score Promedio	Hamming Loss	
Extra Trees	n_estimators=100, max_depth=None, min_samples_split=2, ( $\bar{x} = 0.430$ ) min_samples_leaf=1	0.524, 0.336	0.339, 0.215	0.57	
Random Forest	n_estimators=100, max_depth=None, min_samples_split=2, ( $\bar{x} = 0.4265$ ) min_samples_leaf=1	0.514, 0.339	0.321, 0.203	0.574	
Extra Trees (Optimizado)	n_estimators=300, max_depth=None, min_samples_split=5, min_samples_leaf=2	0.4459	0.2814	0.5541	
Random Forest (Optimizado)	n_estimators=200, max_depth=None, min_samples_split=2, min_samples_leaf=4	0.4434	0.2619	0.5566	

```
MultiOutputClassifier
MultiOutputClassifier(estimator=RandomForestClassifier(random_state=42))
|   estimator: RandomForestClassifier
|   RandomForestClassifier(random_state=42)
|   |
|   * RandomForestClassifier
|       RandomForestClassifier(random_state=42)
```

# MODELO HÍBRIDO DE APRENDIZAJE PROFUNDO PARA CLASIFICACIÓN MULTI-ETIQUETA

- ▶ Combinación de modelos de ensemble y red neuronal
- ▶ Clasificación multi-etiqueta
- ▶ Modelos de Ensemble: Extra Trees (ET) / Random Forest (RF):
- ▶ Arquitectura de Red Neuronal:

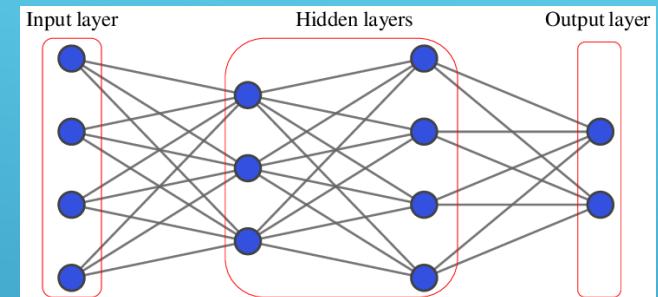
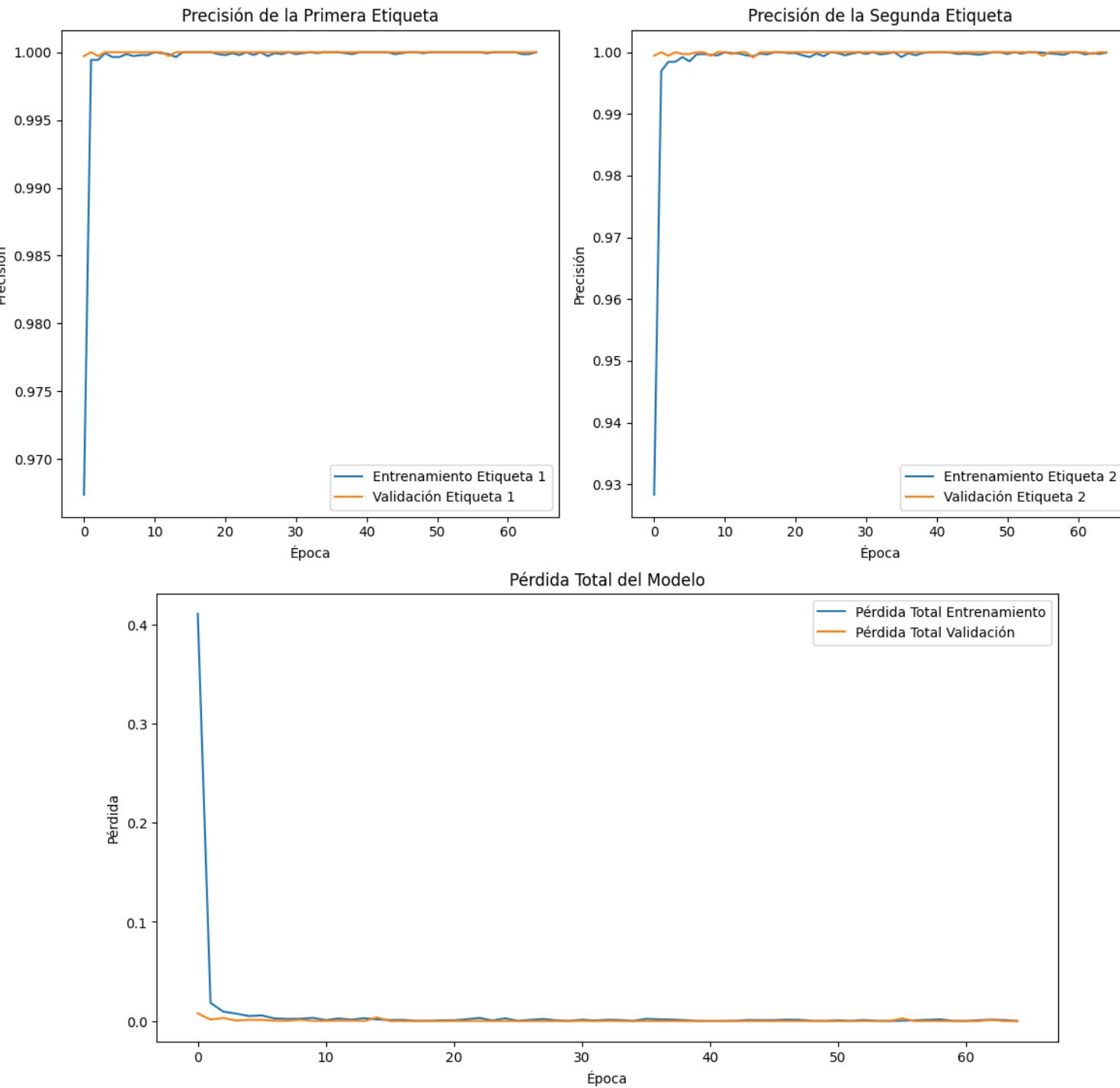


Tabla 3. Tabla de Comparación entre los Modelos de Arquitectura de Red Neuronal

	<b>Modelo 1</b>	<b>Modelo 2 (Mejorado)</b>
<b>Capas compartidas</b>	<ul style="list-style-type: none"><li>• Dense (256 unidades, ReLU)</li><li>• Dropout (30%)</li><li>• Dense (128 unidades, ReLU)</li><li>• Dropout (30%)</li></ul>	<ul style="list-style-type: none"><li>• Dense (128 unidades, ReLU, L2: 0.01)</li><li>• Dropout (40%)</li><li>• Dense (64 unidades, ReLU, L2: 0.01)</li><li>• Dropout (40%)</li></ul>
<b>Rama 1</b>	Dense (softmax)	Dense (softmax, L2: 0.01)
<b>Rama 2</b>	<ul style="list-style-type: none"><li>• Dense (64 unidades, ReLU)</li><li>• Dense (softmax)</li></ul>	<ul style="list-style-type: none"><li>• Dense (32 unidades, ReLU, L2: 0.01)</li><li>• Dropout (40%)</li><li>• Dense (softmax, L2: 0.01)</li></ul>
<b>Optimizador</b>	Adam (lr: 0.001)	Adam (lr inicial: 0.0005)
<b>Épocas</b>	100	100
<b>Batch size</b>	32	32
<b>Early stopping</b>	Paciencia 10	Paciencia 15
<b>Reducción de lr</b>	-	Factor 0.2, paciencia 5

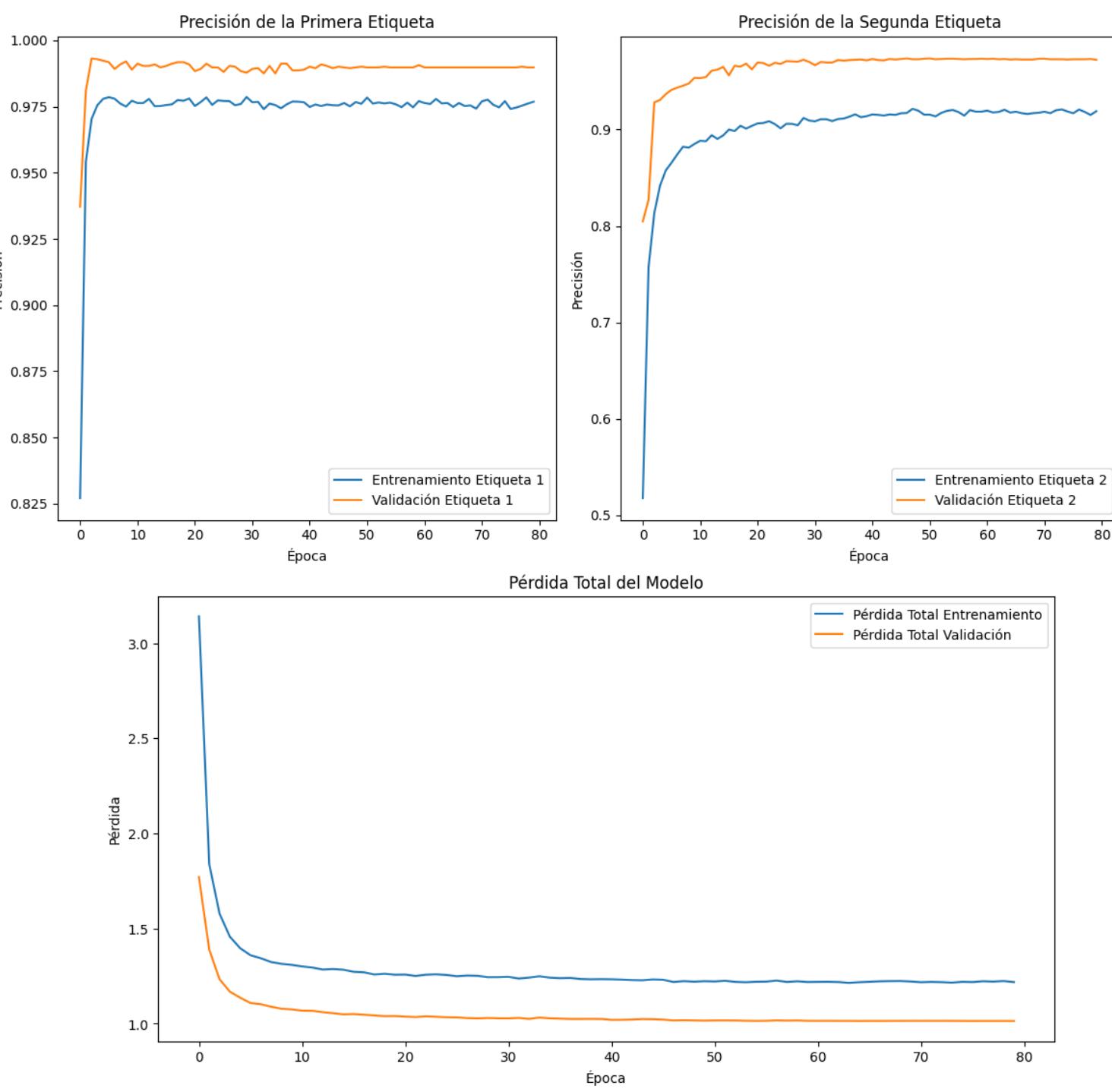
# MODELO 1: SOBREAJUSTE

Sample Footer Text

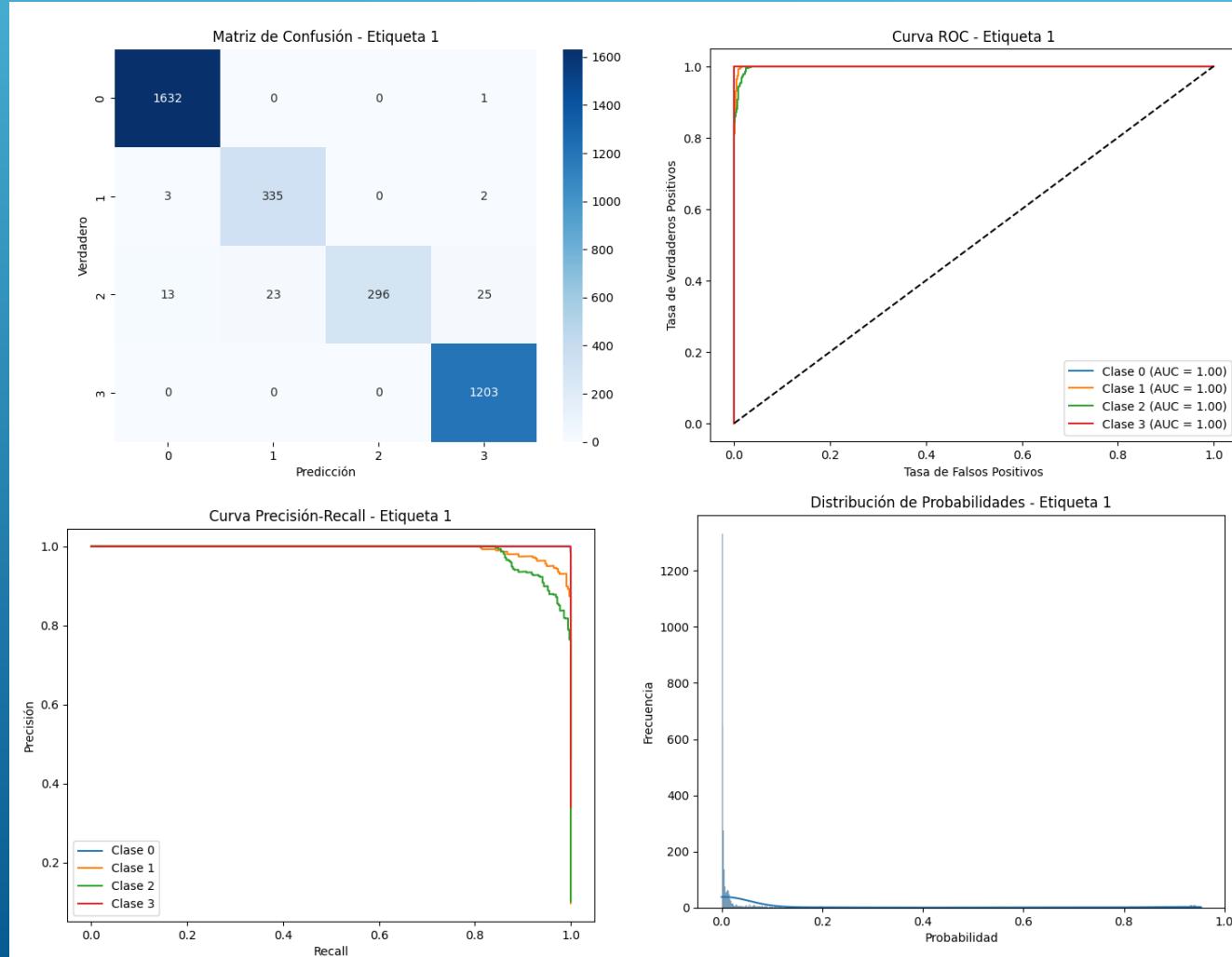


# MODELO 2: MENOS ES MAS

Sample Footer Text

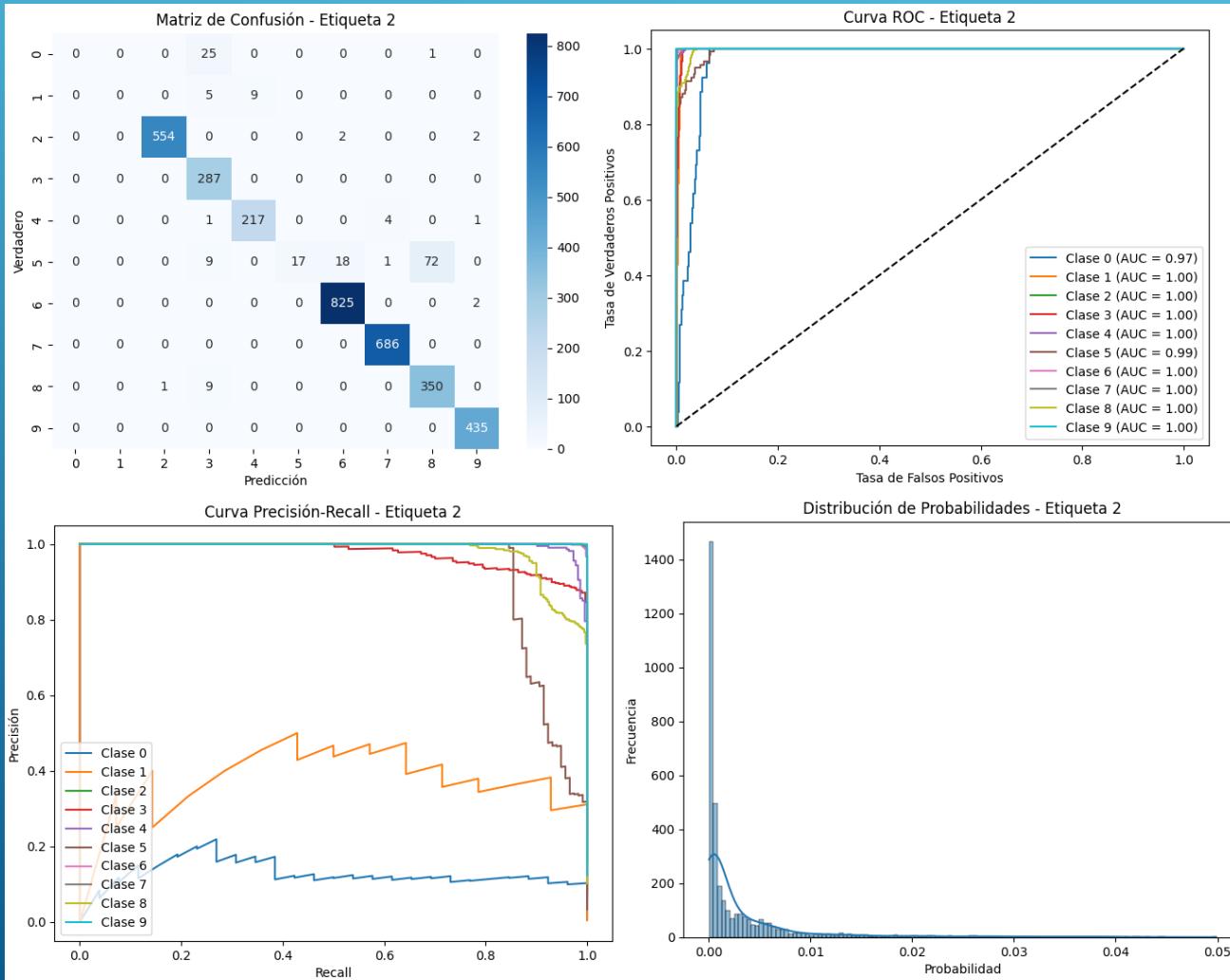


# RENDIMIENTO DEL MODELO PARA LA ETIQUETA 1: DESORDEN GENÉTICO



- ▶ Muestra una precisión casi perfecta, con 1632 y 1203 predicciones correctas para las clases 0 y 3 respectivamente.
- ▶ Todas las clases alcanzan un AUC de 1.00, indicando una discriminación perfecta.
- ▶ Muestra picos de frecuencia extremadamente altos para probabilidades cercanas a 0, con más de 1200 casos. Esto indica que el modelo hace predicciones con un alto grado de certeza.
- ▶ En la Etiqueta 1, todas las clases mantienen una precisión de 1.0 hasta un recall muy alto.

# RENDIMIENTO DEL MODELO PARA LA ETIQUETA 2: ENFERMEDAD

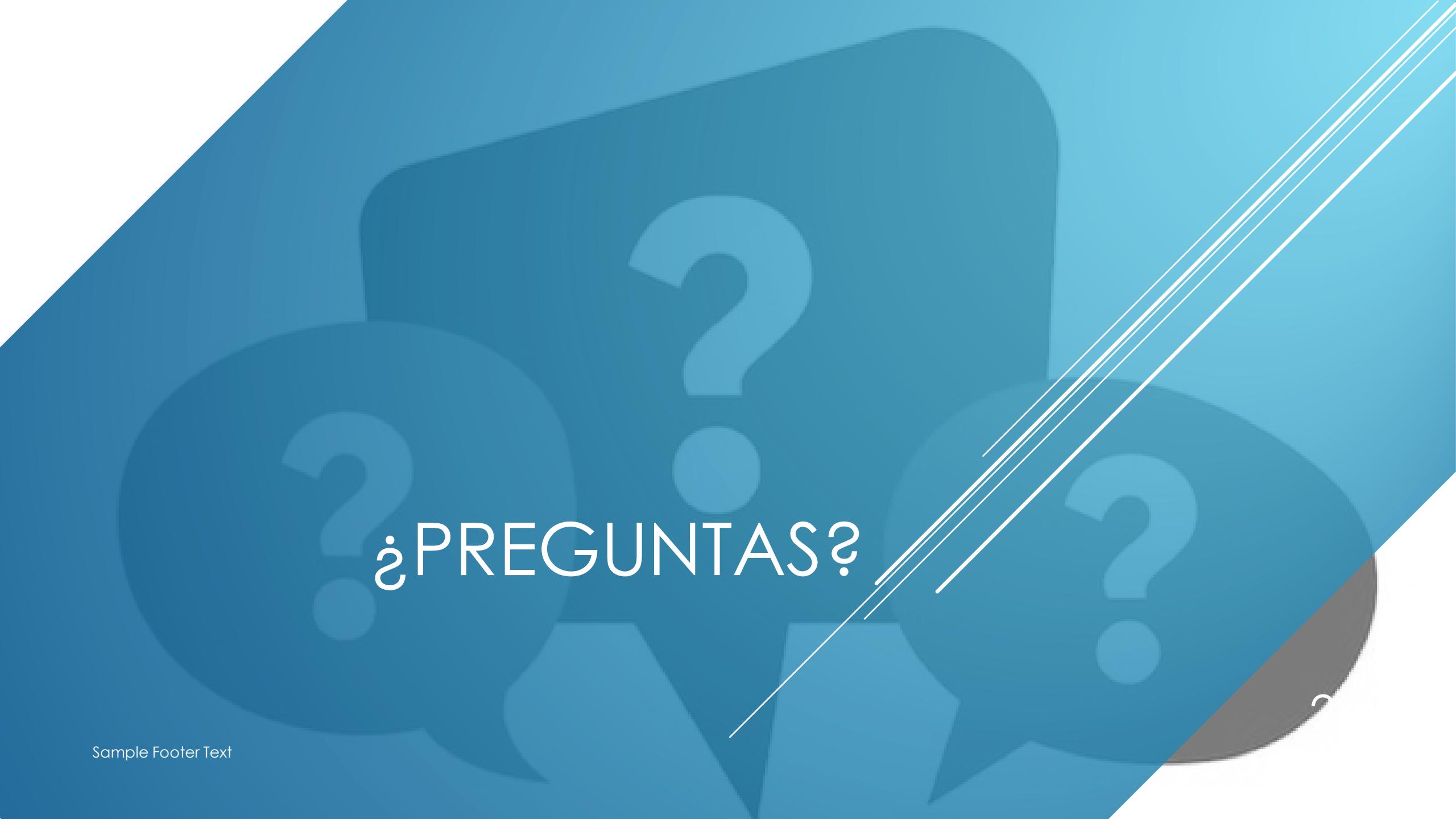


- ▶ Las subclases como la 6 y la 2 tienen 825 y 554 predicciones correctas, demostrando alta precisión.
- ▶ 8 de 10 subclases logran un AUC de 1.00, con la clase más baja alcanzando 0.97.
- ▶ Muestra picos de frecuencia extremadamente altos para probabilidades cercanas a 0, con cerca de 1400 en la Etiqueta 2. Esto indica que el modelo hace predicciones con un alto grado de certeza.
- ▶ La Etiqueta 2 muestra un comportamiento similar a la Etiqueta 1, con solo ligeras caídas en precisión para algunas subclases en niveles muy altos de recall.
- ▶ El modelo mantiene un alto rendimiento en la clasificación más compleja de las 10 subclases.
- ▶ Sin embargo, hay pequeñas confusiones y margen de mejora.



# CONCLUSIONES

- El método desarrollado combina modelos de ensemble (Extra Trees y Random Forest) con una red neuronal profunda (FFNN), aprovechando las fortalezas de ambos enfoques.
- La red está diseñada para predecir secuencialmente dos etiquetas diferentes, lo que permite un aprendizaje más eficiente y potencialmente más robusto.
- La arquitectura utiliza la salida de la primera etiqueta como entrada para la predicción de la segunda, permitiendo que el modelo aprenda relaciones entre las dos tareas de clasificación.
- El modelo maneja eficazmente tanto la clasificación multiclase de la del Desorden Genético (4 clases) como la clasificación más compleja de las Enfermedades (10 clases).
- La alta confiabilidad de las predicciones y la exitosa generalización sugieren que el modelo tiene un gran potencial para aplicaciones prácticas en el mundo real donde se requiera alta precisión en la clasificación de los desórdenes genéticos.
- El uso de modelos de ensemble para generar características iniciales, seguido de una red neuronal relativamente pequeña, sugiere un buen balance entre rendimiento y eficiencia computacional.



# ¿PREGUNTAS?