Predicción de Desórdenes Genéticos en Niños: Un Enfoque de Machine Learning

Resumen del proyecto

Este proyecto se enfoca en desarrollar un modelo de aprendizaje automático para predecir desórdenes genéticos y sus subclases en niños, utilizando un conjunto de datos médicos de alta calidad. El objetivo principal es crear un modelo capaz de predecir con precisión estos desórdenes, lo que podría facilitar el diagnóstico y tratamiento temprano de estas enfermedades. Para ello, se tomará como referencia el enfoque ETRF descrito por Raza et al. (2022)[1], y se construirá una Red Neuronal Feed-Forward Multisalida (FFNN). No se incluirá el análisis ni la revisión de otros modelos.

Antecedentes

Los trastornos genéticos resultan de mutaciones en el genoma o cambios en la estructura de los genes, lo que puede alterar la estructura o función de un organismo. Los genes, compuestos por ácido desoxirribonucleico (ADN), pueden experimentar alteraciones en la secuencia del ADN que conduzcan a trastornos genéticos. La información contenida en los datos genómicos es valiosa para analizar los trastornos que causan enfermedades.

Existen diversos tipos de trastornos genéticos, como los de herencia monogénica, los trastornos cromosómicos, los trastornos mitocondriales y los trastornos complejos o de herencia multifactorial. Todos ellos se estudian en función de la estructura del ADN. Los genes en el ADN contienen información crucial que explica la formación de proteínas. Alteraciones en la estructura del gen pueden dar lugar a proteínas anormales, que no funcionan correctamente en la célula, lo que puede llevar a trastornos como cáncer, diabetes o Alzheimer.

Debido a la complejidad y volumen de los datos genómicos, la predicción manual es laboriosa, propensa a errores e ineficiente. Recientemente, los modelos basados en aprendizaje automático han logrado grandes éxitos en diversas áreas, como el pronóstico, la predicción, la medicina y la automatización.

Este proyecto busca mejorar las capacidades predictivas del modelo descrito por Raza et al. (2022) y presenta las siguientes contribuciones clave:

- Se realizará un análisis exploratorio de datos genéticos (GEDA) para descubrir información valiosa y tendencias en relación con distintos trastornos genéticos.
- Se diseñará un enfoque para extraer características de los datos genómicos utilizando dos modelos de aprendizaje automático: Extra Trees (ET) y Random Forest (RF). Las probabilidades de clase extraídas de los datos de entrenamiento y prueba se combinarán para crear un conjunto de características híbridas, que luego se utilizará para entrenar una red FFNN con el fin de mejorar la precisión de las predicciones.
- A diferencia del enfoque basado en clasificadores en cadena descrito en Raza et al. (2022), este proyecto emplea una FFNN entrenada con las probabilidades de clase de los modelos ET y RF como características. Cada modelo en la red predice en una secuencia específica y utiliza las predicciones de los modelos precedentes para mejorar su precisión.
- Se realizarán experimentos extensos para analizar la precisión, recuperación, exactitud y puntuación F1 del modelo. Además, se comparará su rendimiento en precisión macro, pérdida de Hamming y una puntuación de evaluación.

Metodología

Para este proyecto, se utilizó un conjunto de datos genómicos y genéticos de múltiples etiquetas y clases[2]. Se aplicó un Análisis Exploratorio de Datos Genéticos (GEDA) para identificar factores que puedan causar desórdenes genéticos y

[1] Raza A, Rustam F, Siddiqui HUR, Diez I de la T, Garcia-Zapirain B, Lee E, Ashraf I. 2022. Predicting Genetic Disorder and Types of Disorder Using Chain Classifier Approach. Genes. 14(1):71. doi:https://doi.org/10.3390/genes14010071

[2] Of Genomes And Genetics: HackerEarth Machine Learning Challenge. Obtenido de: https://www.kaggle.com/datasets/aryarishabh/of-genomes-and-genetics-hackerearth-ml-challenge/data.

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

Figura 1. Detalle de los atributos del conjunto de datos: Cada fila de la tabla corresponde a un atributo diferente y proporciona una descripción que explica qué tipo de datos tiene cada atributo en el conjunto de datos.

obtener información útil sobre el conjunto de datos. Posteriormente, se emplearon técnicas de ingeniería de características para mapear los datos y seleccionar las características más relevantes, con el objetivo de mejorar el rendimiento de los modelos. Además, se equilibró la distribución de las clases de desórdenes genéticos en los datos para entrenar un modelo de aprendizaje con un número igual de muestras. Finalmente, se implementó la técnica de extracción de características ETRF, que enriquece el conjunto de características utilizadas para entrenar todos los modelos.

Conjunto de Datos Genómicos

El conjunto de datos genómicos y genéticos se basa en información médica de pacientes, tanto niños como adultos, que padecen desórdenes genéticos. Este conjunto de datos es de tipo multi-etiqueta y multi-clase. La primera etiqueta corresponde al "desorden genético" y la segunda sub-etiqueta se refiere a la "subclase de desorden". En total, el conjunto de datos contiene 45 atributos que se detallan en la Figura 1. En la Tabla 1 se presenta un resumen estadístico de las variables numéricas que proporcionan una visión integral de las características demográficas y biométricas del conjunto de datos.

Análisis Exploratorio de Datos Genéticos (GEDA)

El GEDA se aplicó al conjunto de datos genómicos para descubrir patrones ocultos y obtener información relevante que pudiera ser útil en la predicción de desórdenes genéticos. Este análisis incluyó la generación de gráficos como gráficos de pares, análisis de distribución de datos en 3D, gráficos de barras, gráficos de dispersión y gráficos circulares. El GEDA sirvió como una herramienta valiosa para identificar información estadística significativa a partir de los datos genéticos.

Ingeniería de Características y Normalización de Datos

La ingeniería de características es un proceso crucial para los modelos de aprendizaje automático. Se aplicaron técnicas de ingeniería de características para codificar y mapear los datos del conjunto de datos genómicos. Se determinó la importancia de las características utilizando un modelo de árbol de decisión (DTC), y se eliminaron características con baja o nula importancia (Figura 2). Entre las características eliminadas se encuentran: ‘patient Id’, ‘patient first name’, ‘family

Tabla 1. Resumen estadístico de las variables numéricas relacionadas con los pacientes del conjunto de datos

	Edad del Paciente	Edad de la madre	Edad del padre	Numero de abortos previos	Conteo de globulos rojos(millones/µL)	Conteo de globulos blancos (K/µL)
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

name', 'father's name', 'institute name', 'location of institute', 'place of birth', y 'parental consent', entre otras, ya que no contribuyeron significativamente a la predicción de desórdenes genéticos. En total, se eliminaron 13 características.

Los valores nulos en el conjunto de datos fueron reemplazados por ceros. Las características seleccionadas se codificaron adecuadamente con valores categóricos, asignando valores numéricos a opciones como "Sí" y "No", así como a otras categorías relevantes.

División de Datos

La división de datos se aplica para separar el conjunto de datos en un conjunto de entrenamiento y un conjunto de prueba. Esta división ayuda a reducir el sobreajuste del modelo y a mejorar su capacidad de generalización. En este proyecto, se utilizó una división estándar de 80:20, es decir, el 80% de los datos se destinó al entrenamiento del modelo, y el 20% restante se utilizó para su evaluación. Esta división fue seleccionada por su efectividad probada en la construcción de modelos de aprendizaje.

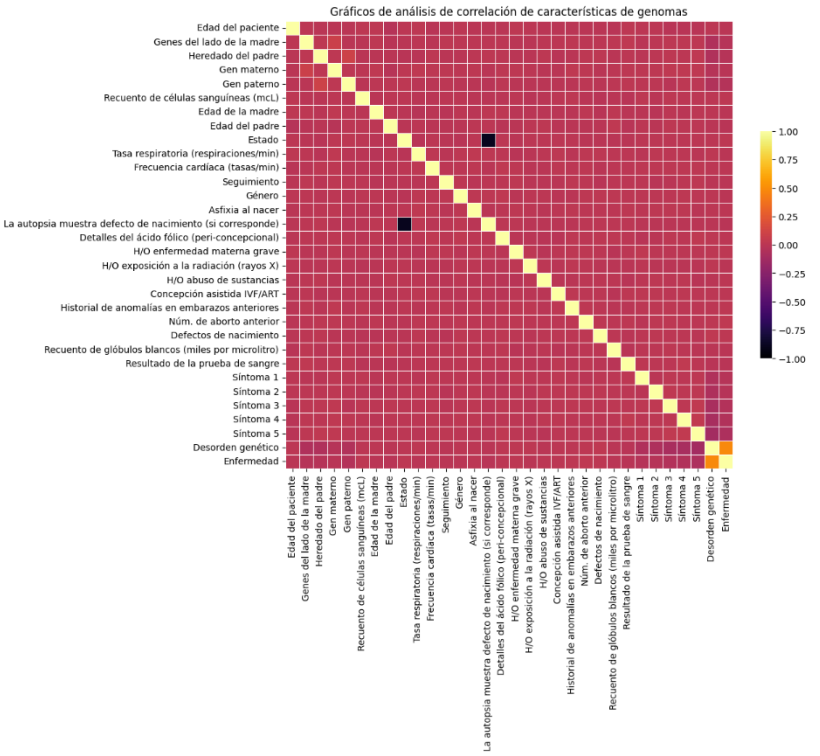

Figura 2. ***Analisis de correlación del conjunto de datos genéticos:*** *Cada celda del mapa de calor representa la correlación entre dos características. Los colores representan el grado de correlación entre ellas, variando desde el amarillo claro (correlación positiva) hasta el negro (correlación negativa).*

Técnicas de Aprendizaje Aplicadas

Se aplicaron varios modelos de aprendizaje automático para analizar el rendimiento del enfoque propuesto de ingeniería de características. A continuación se presenta una descripción breve de cada uno de estos modelos en términos de su arquitectura y mecanismos de funcionamiento.

- RFC (Random Forest Classifier) es un modelo supervisado basado en múltiples árboles de decisión. Toma predicciones de varios árboles, y la predicción final se selecciona mediante votación mayoritaria, un enfoque conocido como bagging. RFC selecciona aleatoriamente observaciones para construir los árboles de decisión, lo que reduce los problemas de sobreajuste y mejora el rendimiento de clasificación.
- ETC (Extra Trees Classifier) es otra técnica de ensamblado basada en árboles de decisión. Es similar a RFC, pero se diferencia en que ETC sigue una selección aleatoria de divisiones en los valores, lo que reduce la varianza del modelo. Esto da como resultado una mayor precisión y menor sobreajuste.
- MLP (Multi-Layer Perceptron) es un algoritmo de clasificación que utiliza una red neuronal feedforward. Consiste en múltiples capas totalmente conectadas, y su proceso de entrenamiento es iterativo, utilizando descenso de gradiente estocástico para optimizar la función de pérdida. MLP ha demostrado ser superior para varias tareas en comparación con otros modelos más complejos.
- Se utilizó KNN (K-Nearest Neighbors) exclusivamente para imputar características problemáticas durante el proceso de ingeniería de características. KNN es una técnica supervisada no paramétrica que agrupa los datos en función de la proximidad a sus vecinos más cercanos. No se utilizó KNN como parte de los modelos de aprendizaje aplicados en el experimento.
- El algoritmo DTC (Decision Tree Classifier) se utilizó para determinar las características con baja o nula importancia. Es una técnica de aprendizaje supervisado utilizada para tareas de clasificación. DTC se organiza en una estructura similar a un árbol, con nodos y hojas. Los nodos internos contienen los atributos de los datos y realizan divisiones basadas en ellos, mientras que las etiquetas de resultado se encuentran en los nodos hoja. El nodo raíz de un DTC es el más alto en la jerarquía. Los algoritmos DTC construyen árboles de decisión automáticamente a partir de los datos de entrada. El objetivo principal de la técnica DTC es identificar el árbol de decisión óptimo

minimizando el error de generalización. Uno de los principales desafíos en DTC es la selección de atributos de los datos.

Enfoque de Ingeniería de Características ETRF

El enfoque ETRF descrito en Raza et al., 2022 combina los algoritmos ET y RF para su uso como técnica de extracción de características en la construcción de modelos de aprendizaje y la predicción de desórdenes genéticos. La formación y el mecanismo de extracción del conjunto de características del conjunto de datos genómicos utilizando la técnica ETRF se visualizan en la Figura 3.

El análisis arquitectónico muestra cómo se introducen los datos genómicos en los algoritmos ET y RF por separado, y las probabilidades de clase predichas se extraen de ambas técnicas. Luego, se forma un conjunto de características híbrido combinando las probabilidades de clase extraídas, que posteriormente se utiliza como entrada para las técnicas de aprendizaje aplicadas en la predicción de desórdenes genéticos y sus subtipos.

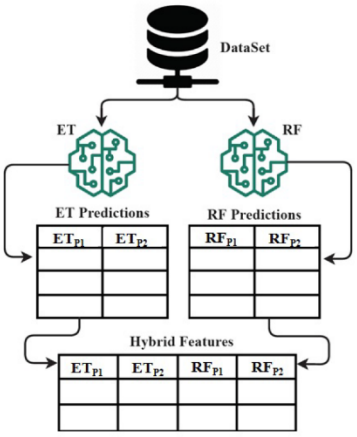

Figura 3. Arquitectura de la técnica ET/RF para la producción de características híbridas: *El conjunto de datos que se divide en dos caminos: uno que lleva a un bloque ET y otro que lleva a un bloque RF. Cada bloque genera sus predicciones. Estas predicciones se utilizan para crear un conjunto de 'Características Híbridas', que recibe entradas de las predicciones tanto de ET como de RF.*

Configuración Experimental

Los experimentos se realizaron en una CPU Intel i7-7700HQ, con 24 GB de memoria RAM y una tarjeta gráfica NVIDIA GTX1650. Se utilizaron las herramientas de Python 3.12.4 en Visual Studio Code 1.91 para construir los modelos de aprendizaje automático.

Métricas de Evaluación

Se utilizaron varias métricas de evaluación para medir el rendimiento de los modelos ETRF, incluyendo la precisión macro, sensibilidad, precisión, pérdida de Hamming y la puntuación F1. A continuación se describen brevemente estos factores:

- Verdadero Positivo (TP): Número de muestras positivas correctamente clasificadas por el modelo.
- Verdadero Negativo (TN): Número de muestras negativas correctamente clasificadas por el modelo.
- Falso Positivo (FP): Número de muestras negativas incorrectamente clasificadas como positivas por el modelo.
- Falso Negativo (FN): Número de muestras positivas incorrectamente clasificadas como negativas por el modelo.

Para problemas de etiquetas múltiples, las métricas basadas en etiquetas se evalúan para cada etiqueta y luego se promedian. La precisión macro se calcula en cada clase individual y luego se promedia en todas las clases. La pérdida de Hamming calcula la proporción de etiquetas objetivo incorrectamente predichas respecto al total de etiquetas.

Además, se calcularon las métricas de precisión, sensibilidad y F1 para medir el rendimiento de los modelos, donde la puntuación F1 es la media armónica de precisión y sensibilidad.

Red Neuronal para Predicción de Desórdenes Genéticos

Tabla 2. Comparación de Modelos de Clasificación Multi-Salida Extra Trees/Random Forest*: Se presenta la evaluación comparativa de las diferentes configuraciones de los modelos de aprendizaje.*

Modelo	Parámetros	Exactitud Promedio	F1-Score Promedio	Hamming Loss
Extra Trees	n_estimators=100, max_depth=None, min_samples_split=2, min_samples_leaf=1	0.524, 0.336 ($\bar{x}$ = 0.430)	0.339, 0.215	0.57
Random Forest	n_estimators=100, max_depth=None, min_samples_split=2, min_samples_leaf=1	0.514, 0.339 ($\bar{x}$ =0.4265)	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

La red neuronal diseñada en este proyecto se basa en la salida de probabilidades generadas por dos modelos: un Extra Trees (ET) y un Random Forest (RF), ambos configurados para tareas de clasificación múltiple. Esta combinación de modelos se utiliza para crear un conjunto de características híbrido que alimenta la red neuronal, cuyo objetivo es predecir dos etiquetas distintas relacionadas con desórdenes genéticos en niños. Se entrenaron dos modelos de árboles de decisión, ET y RF, utilizando un conjunto de datos de entrenamiento limpio. Los parámetros de los modelos se muestran en la Tabla 2.

Tabla 3. Arquitectura de los modelos de Red Neuronal evaluados: *Se presenta la comparación de las diferentes arquitecturas de redes neuronales. Cada modelo se evalúo con base a varios parámetros.*

Modelo	Capas Compartidas	Rama para Output1	Rama para Output2	Optimización	Pérdida	Tasa de Aprendizaje	Callbacks
Modelo 1	- Dense (256, ReLU) - Dropout (0.3) - Dense (128, ReLU) - Dropout (0.3)	- Dense (n_classes_1, Softmax)	- Dense (64, ReLU) - Dense (n_classes_2, Softmax)	Adam	Categorical Crossentropy	0.001	EarlyStopping (patience=10)
Modelo 2	- Dense (128, ReLU) - Dropout (0.3) - Dense (64, ReLU) - Dropout (0.3)	- Dense (n_classes_1, Softmax)	- Dense (32, ReLU) - Dense (n_classes_2, Softmax)	Adam	Categorical Crossentropy	0.001	EarlyStopping (patience=10)
Modelo 3	- Dense (128, ReLU, L2) - Dropout (0.4) - Dense (64, ReLU, L2) - Dropout (0.4)	- Dense (n_classes_1, Softmax, L2)	- Dense (32, ReLU, L2) - Dropout (0.4) - Dense (n_classes_2, Softmax, L2)	Adam	Categorical Crossentropy	0.0005	EarlyStopping (patience=15) ReduceLROnPlateau (factor=0.2, patience=5, min_lr=0.00001)
Modelo 4	- Dense (64, ReLU, L2) - Dropout (0.4) - Dense (32, ReLU, L2) - Dropout (0.4)	- Dense (n_classes_1, Softmax, L2)	- Dense (16, ReLU, L2) - Dropout (0.4) - Dense (n_classes_2, Softmax, L2)	Adam	Categorical Crossentropy	0.0005	EarlyStopping (patience=15) ReduceLROnPlateau (factor=0.2, patience=5, min_lr=0.00001)

Arquitectura del Modelo

Para la etapa de modelado final, se implementó una red neuronal que utiliza las probabilidades de clase generadas por los modelos ET y RF como entrada. Este enfoque se diseñó para mejorar la precisión en la clasificación de los desórdenes genéticos y sus subclases, integrando características provenientes de ambos modelos en un conjunto de características híbrido.

Preprocesamiento de Etiquetas

Las etiquetas de los desórdenes genéticos y las subclases asociadas se codificaron utilizando *LabelEncoder* y posteriormente convertidas a una representación *one-hot encoding* para ser compatibles con la salida de la red neuronal.

Definición de la Arquitectura

La red neuronal cuenta con una capa de entrada que toma las características híbridas, las cuales combinan las probabilidades de clase de los modelos ET y RF.

A partir de esta entrada, se construyen capas densas con activación ReLU y Dropout que actúan como reguladores. Estas capas extraen patrones generales de las características híbridas y son compartidas entre las dos ramas de la red.

La arquitectura está dividida en dos ramas: la primera rama tiene su propia capa densa y predice la clase principal del desorden genético (*output1*). La salida de esta primera etiqueta se concatena con las características procesadas, y se utiliza como entrada para la siguiente rama de la red. La segunda rama toma esta predicción junto con las características originales para predecir la segunda etiqueta, correspondiente a la subclase del desorden genético (*output2*). Los detalles de las arquitecturas utilizadas se muestran en la Tabla 3.

Entrenamiento del Modelo

El entrenamiento se llevó a cabo con validación cruzada, utilizando una porción del 20% de los datos de entrenamiento como conjunto de validación, durante un máximo de 100 épocas. Se incluyó la técnica de *EarlyStopping* para detener el entrenamiento si no se observaban mejoras en la validación.

Evaluación del Modelo

La red neuronal se evaluó en el conjunto de prueba, obteniendo precisiones específicas para cada una de las etiquetas. Las métricas reportadas incluyen la pérdida y la precisión para ambas salidas del modelo. Además, se generaron predicciones sobre los datos de prueba y se transformaron de nuevo a sus etiquetas originales para su interpretación.

Visualización y Análisis de Resultados

Para evaluar y visualizar el rendimiento de la red neuronal, se emplearon diversas técnicas gráficas:

- Curvas de Precisión y Pérdida: Se graficaron las curvas de precisión y pérdida para el entrenamiento y la validación de ambas etiquetas, permitiendo analizar el comportamiento del modelo a lo largo de las épocas.

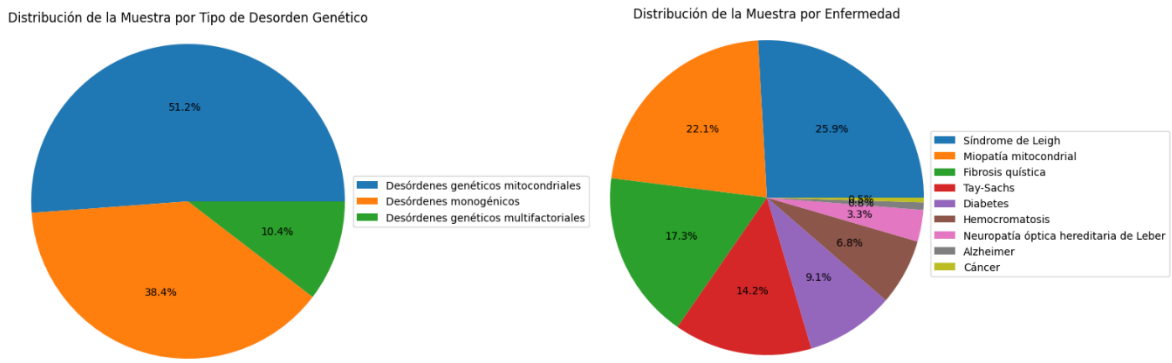

Figura 4. Distribución de las muestras en las diferentes clases del conjunto de datos: *El primer gráfico muestra la distribución de los tipos de desorden genético en la muestra. El segundo gráfico muestra la distribución de la muestra por enfermedad.*

- Matriz de Confusión: Se generaron matrices de confusión para cada una de las etiquetas, lo cual facilita la comprensión de la capacidad del modelo para distinguir entre las diferentes clases.
- Curvas ROC y Curvas de Precisión-Sensibilidad: Estas curvas se utilizaron para evaluar el rendimiento del modelo en un contexto de clasificación multiclase, calculando el área bajo la curva (AUC) para cada clase.
- Distribución de Probabilidades: Se analizaron las distribuciones de las probabilidades predichas, proporcionando información sobre la confianza del modelo en sus predicciones.

Estos análisis permiten validar la capacidad predictiva de la red neuronal en la identificación de desórdenes genéticos y sus subclases, y proporcionan un marco robusto para futuras mejoras en la precisión del modelo.

Resultados

Distribución de Etiquetas de Desórdenes Genéticos y Subclases

El análisis de la distribución de datos de las etiquetas de desórdenes genéticos y subclases (Figura 4) reveló que la clase de desórdenes de herencia genética mitocondrial presenta la mayor cantidad de muestras, mientras que los desórdenes de herencia genética multifactorial tienen la menor cantidad. Entre las subclases de desórdenes, las enfermedades de neuropatía óptica hereditaria de Leber y diabetes muestran los valores de distribución más bajos, mientras que Tay-Sachs también tiene una representación relativamente baja.

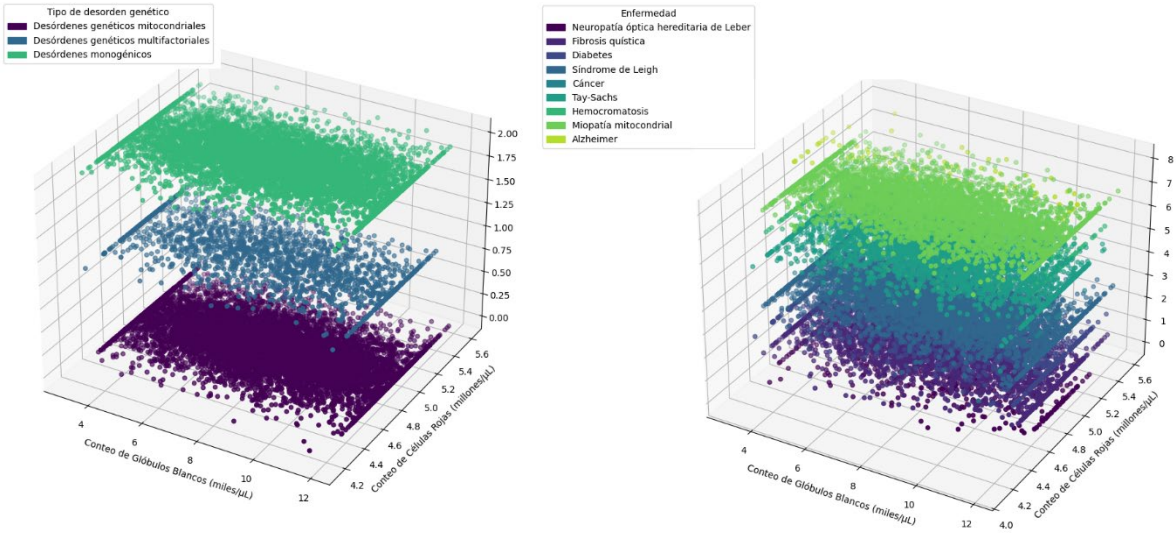

Figura 5. Analisis de dispersión 3D para el recuento de glóbulos blancos y el recuento de células sanguíneas por categoría de desorden genético (izquierda) y enfermedad (derecha)*: Cada punto en el gráfico representa una muestra individual, y su posición en el espacio tridimensional corresponde a sus valores de recuento de glóbulos blancos y células sanguíneas. Los colores de los puntos indican las diferentes categorías de desorden genético y enfermedad.*

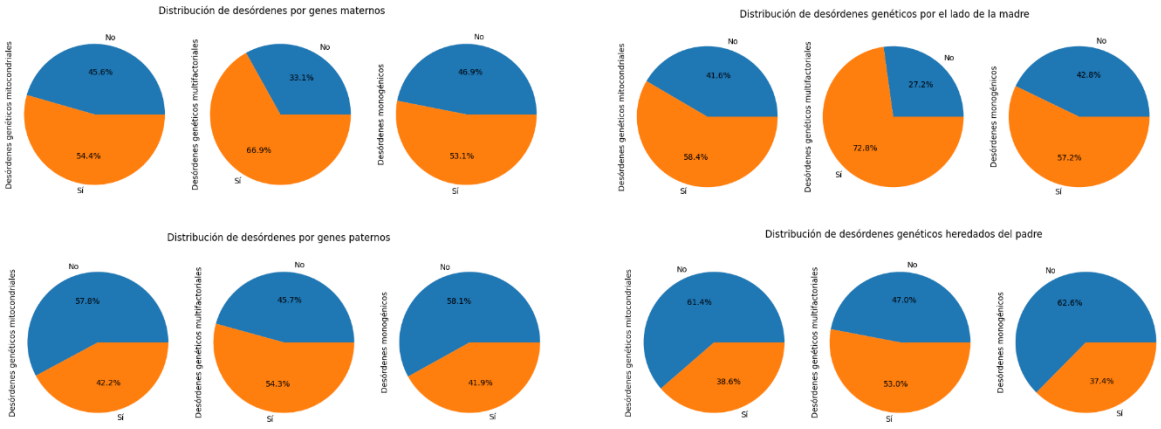

Figura 6. Distribución de los genomas por tipo de desorden genético: *Se presentan doce gráficos circulares que muestran si el desorden genético es heredado del padre o la madre. Los gráficos están divididos en dos categorías principales: desórdenes heredados por vía materna (superior) y desórdenes heredados por vía paterna (inferior). Cada categoría se subdivide a su vez en dos subcategorías basadas en el tipo de herencia, los desórdenes genéticos producidos por alteraciones en el material genético que pueden ocurrir de manera espontánea durante el desarrollo embrionario y pueden o no heredarse (izquierda), y desórdenes hereditarios, que son aquellos transmitidos de padres a hijos de generación en generación (derecha).*

Análisis de Distribución 3D de Conteo de Células Sanguíneas

Se llevó a cabo un análisis de la distribución de datos en 3D del conteo de glóbulos blancos (miles por microlitro) y el conteo de células sanguíneas (mcL) con la etiqueta de desorden genético (Figura 5). Se observó que cuando el conteo de glóbulos blancos es menor a 0, se encuentra un desorden genético de cualquier tipo. Sin embargo, cuando el conteo de glóbulos blancos está entre 0 y 2, no hay posibilidad de desorden genético mitocondrial, aunque se encuentran desórdenes multifactoriales y de un solo gen. Los valores de conteo de células sanguíneas menores a 4.2 mcL indican la ausencia de desórdenes genéticos.

El análisis también se extendió a las subclases de desórdenes genéticos, donde se encontró que no existen subclases de enfermedades cuando el valor del conteo de células sanguíneas varía entre 4.2 y 4.4 mcL. La neuropatía óptica hereditaria

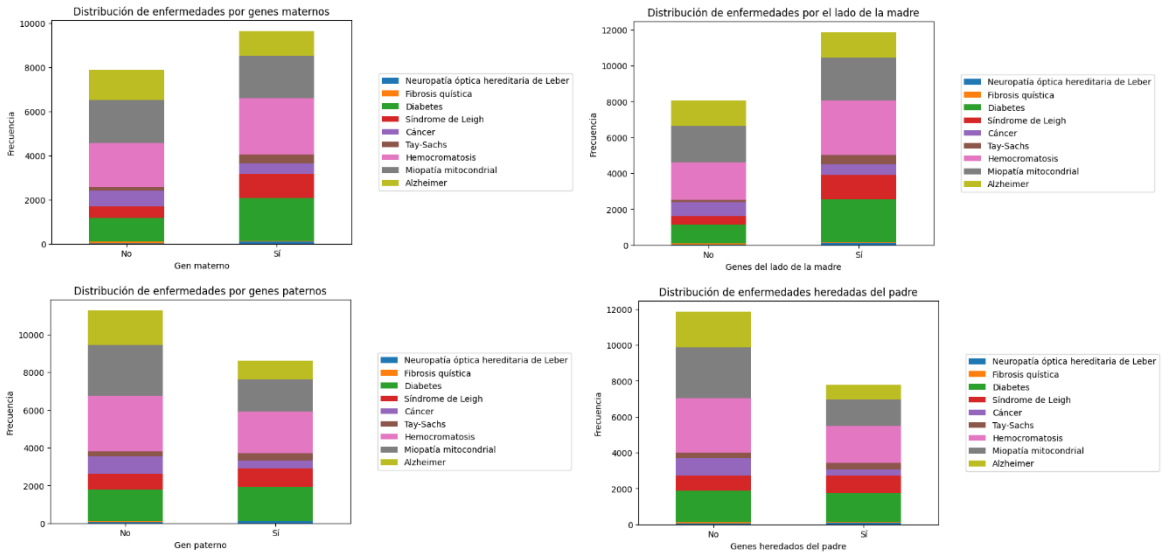

Figura 7. Distribución de genomas por enfermedad: *Se presentan ocho gráficos de barras apiladas que muestran la proporción de enfermedades genéticas en el conjunto de datos. Los gráficos se organizan en dos categorías principales: desórdenes transmitidos por línea materna (superior) y desórdenes transmitidos por línea paterna (inferior). A su vez, cada categoría se subdivide en dos tipos de herencia, diferenciando entre desórdenes genéticos, que pueden surgir espontáneamente y podrían o no ser heredados (izquierda), y desórdenes hereditarios, que son aquellos que se transmiten de una generación a otra (derecha).*

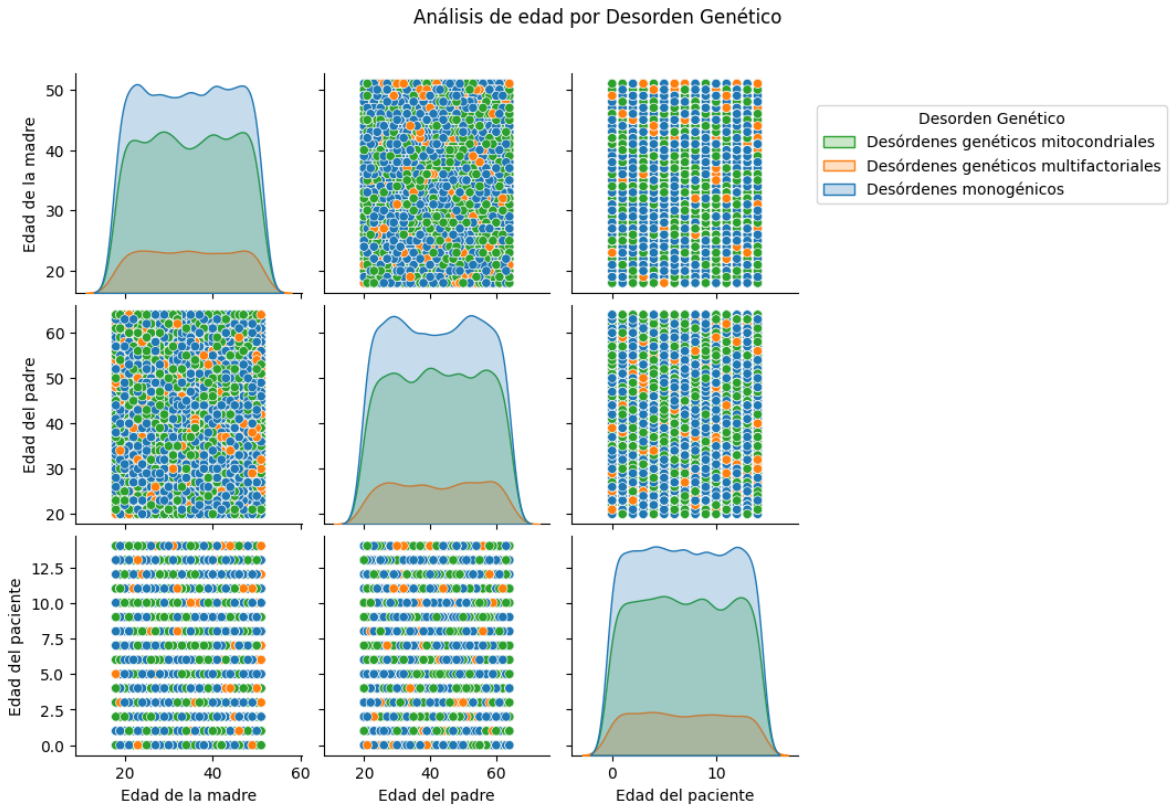

Figura 8. Analisis de la edad de los pacientes para los desórdenes genéticos*: Se muestran los gráficos de dispersión con puntos de datos individuales, comparando la edad de la madre, la edad del padre y la edad del paciente en relación con los diferentes tipos de desórdenes genéticos, además en la diagonal se muestra un gráfico de densidad que representa la concentración de casos correspondiente a las edades.*

de Leber se encuentra cuando el conteo de células sanguíneas varía entre 4.4 y 4.8 mcL, y un valor de conteo de células sanguíneas superior a 4.8 mcL indica la ocurrencia de todas las subclases de desórdenes.

Análisis de Genes Heredados

Se analizó el impacto de los genes heredados en la probabilidad de desórdenes genéticos (Figura 6). Los resultados mostraron que, cuando el valor del gen materno es positivo, la probabilidad de desorden mitocondrial es mayor, mientras que la probabilidad de desorden de un solo gen es menor. Para las subclases de desórdenes genéticos (Figura 7), se observó que el desorden de diabetes tiene una alta ocurrencia sin importar el origen del gen, mientras que el síndrome de Tay-Sachs tiene bajas probabilidades de ocurrir en todos los genes analizados.

Análisis del Factor Edad en Desórdenes Genéticos

Se analizó el impacto de la edad de la madre, el padre y el paciente en la probabilidad de desórdenes genéticos (Figura 8). Los resultados indicaron que hay altas probabilidades de desórdenes genéticos cuando la edad de la madre está entre 20 y 60 años, y cuando la edad del padre está entre 20 y 70 años. Además, los desórdenes genéticos tienden a manifestarse antes de los 15 años en los pacientes.

Análisis Comparativo de Modelos

Durante el proyecto, se desarrollaron cuatro modelos de redes neuronales (ver Tabla 3). Inicialmente, se diseñó un modelo de alta complejidad que resultó en un notable sobreajuste. Este primer modelo (No se muestra) tuvo un rendimiento excepcional de entre 0.995 y 1.000 a lo largo de las épocas para predecir el desorden genético (Etiqueta 1), lo que refleja una precisión extremadamente alta en ambos conjuntos de datos, pero su capacidad para generalizar en el conjunto de validación fue significativamente menor, lo que sugiere que el modelo capturó ruidos específicos del conjunto de entrenamiento en lugar de patrones generales. La consistencia en las precisiones entre los conjuntos de entrenamiento y validación sugiere que hay un sobreajuste significativo para esta etiqueta, indicando que el modelo no generaliza bien en

los datos de prueba. Aún más, para la predicción de enfermedades (Etiqueta 2), se observa que la curva de entrenamiento oscila entre 1.10 y 1.14, mientras que la curva de validación fluctúa entre 0.230 y 0.235. Está marcada diferencia entre las precisiones del entrenamiento y la validación indica la presencia de un sobreajuste significativo. El modelo está logrando un rendimiento mucho mejor en el conjunto de entrenamiento en comparación con el conjunto de validación, lo que sugiere que ha memorizado los datos de entrenamiento sin capturar adecuadamente las relaciones subyacentes que permitan una buena generalización. Finalmente, las curvas de pérdida para el entrenamiento y la validación muestran una tendencia decreciente a lo largo de las épocas, lo que indica una disminución continua en la pérdida total del modelo. Aunque ambas curvas descienden, la persistente diferencia entre la pérdida de entrenamiento y la de validación, junto con la observada en la precisión de la segunda etiqueta, refuerza la sospecha de que el modelo podría estar sobre ajustándose, particularmente en lo que respecta a la segunda etiqueta.

Para abordar el problema de sobreajuste observado en el primer modelo, se desarrollaron sucesivamente tres modelos adicionales, cada uno con una arquitectura menos compleja que el anterior. El proceso de simplificación incluyó la reducción del número de capas, la cantidad de neuronas en cada capa, y la implementación de técnicas de regularización más estrictas.

El cuarto y último modelo, que presenta la estructura más sencilla de todos, demostró ser el más eficaz. Este modelo no solo evitó el sobreajuste, sino que también mantuvo un rendimiento competitivo en ambos conjuntos de datos, de

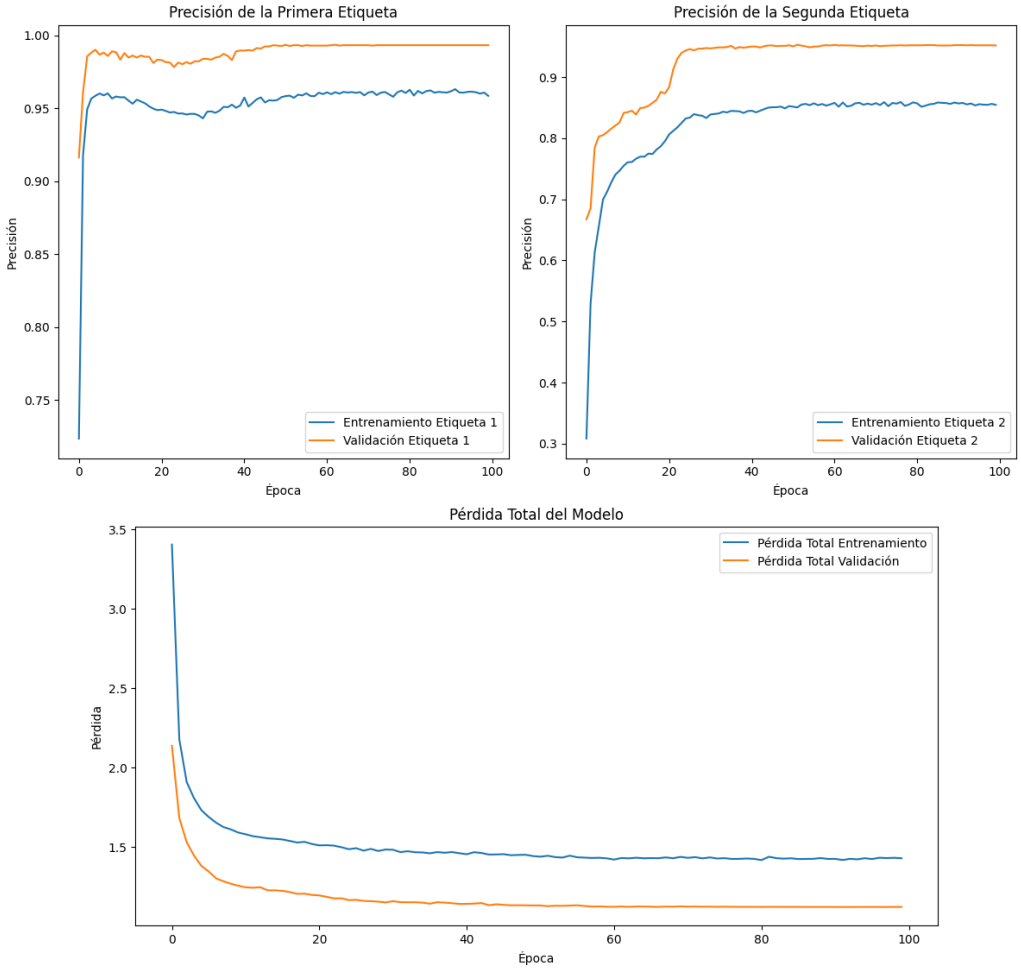

Figura 9. Evaluación de la precisión y pérdida durante el entrenamiento del modelo 4*. Los dos gráficos superiores representan la precisión de la Etiquetas 1 y la Etiqueta 2 a lo largo del tiempo (épocas) para los conjuntos de entrenamiento y validación. El gráfico inferior muestra la pérdida total del modelo durante las épocas para los mismos conjuntos.*

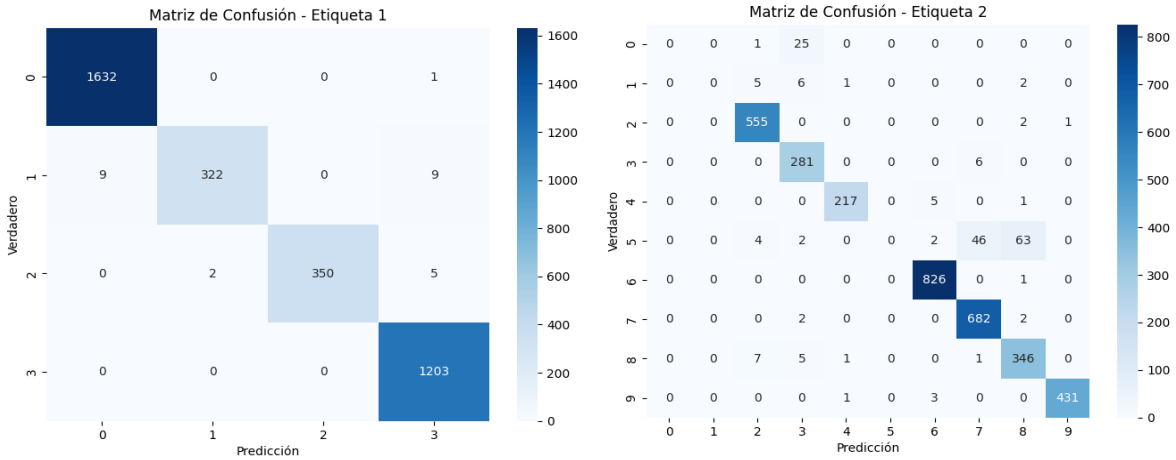

Figura 10. Análisis del desempeño del modelo 4: Matrices de Confusión. *Se presentan las dos matrices de confusión para las Etiquetas 1 y 2, respectivamente. Cada matriz proporciona una representación visual de la precisión del modelo para cada clase mostrando el número de verdaderos positivos (diagonal) en comparación con los falsos positivos y falsos negativos (elementos fuera de la diagonal).*

entrenamiento y validación. Los resultados obtenidos con este modelo indican que una arquitectura menos compleja es más adecuada para el conjunto de datos utilizado en este proyecto, proporcionando un equilibrio óptimo entre precisión y capacidad de generalización.

La evolución de la precisión durante el entrenamiento y la validación para las dos etiquetas se muestran en la Figura 9. Para la primera etiqueta, se observa que la precisión en el conjunto de entrenamiento aumenta rápidamente en las primeras épocas, estabilizándose alrededor de la época 20, alcanzando un valor cercano al 95%. En el conjunto de validación, la precisión es consistentemente superior a la del entrenamiento, logrando valores que superan el 97% tras las primeras 10 épocas y manteniéndose estable. Este comportamiento sugiere que el modelo se ajusta correctamente a los datos de entrenamiento y generaliza bien en el conjunto de validación. La precisión para la segunda etiqueta presenta un comportamiento similar, pero con valores iniciales más bajos y una convergencia más lenta. Sin embargo, la precisión final en la validación supera ligeramente el 90%, lo cual es indicativo de un modelo robusto, aunque posiblemente se deba a la mayor complejidad en la predicción de esta etiqueta en comparación con la primera.

En la Figura 9 también se muestra la evolución de la pérdida total del modelo durante el entrenamiento y la validación. Se observa una rápida disminución de la pérdida en ambas curvas en las primeras 20 épocas, con la pérdida de entrenamiento estabilizándose alrededor de un valor de 1.2, mientras que la pérdida de validación se mantiene alrededor de 1.4. Este comportamiento es consistente con lo observado en las métricas de precisión, sugiriendo que el modelo ha logrado un buen ajuste a los datos sin signos evidentes de sobreajuste. La menor pérdida en el conjunto de validación en comparación con el conjunto de entrenamiento refuerza la idea de una buena capacidad de generalización del modelo.

Análisis de resultados para las dos etiquetas

La matriz de confusión (Figura 10) de la etiqueta 1 muestra una clasificación casi perfecta para todas las clases, con valores altos en la diagonal principal (1632, 322, 350, 1203) y errores mínimos. La curva ROC (Figura 11) alcanza el punto óptimo (0,1) para todas las clases, con un área bajo la curva (AUC) de 1.00, indicando una discriminación perfecta. Las curvas de precisión-sensibilidad (Figura 11) exhiben un rendimiento excepcional, manteniéndose cerca del punto (1,1) para todas las clases, lo que sugiere alta precisión y sensibilidad simultáneos. La distribución de probabilidades (Figura 12) revela una concentración significativa cerca de 0 y 1, con poca dispersión en el medio, indicando que el modelo hace predicciones con alta confianza en la mayoría de los casos.

Para la etiqueta 2 la matriz de confusión (Figura 10) muestra un buen rendimiento general, con la mayoría de las predicciones en la diagonal principal. Las clases 2, 6 y 7 destacan con 555, 826 y 682 aciertos respectivamente, aunque se observan algunas confusiones, notablemente entre las clases 5 y 8. La curva ROC Figura 11) indica un excelente desempeño global, con AUC cercano o igual a 1 para todas las clases, siendo ligeramente menor (0.98-0.99) para las clases 0, 5 y 8. Las curvas de precisión-sensibilidad (Figura 11) muestran un rendimiento variado: las clases 2, 3, 4, 6, 7 y

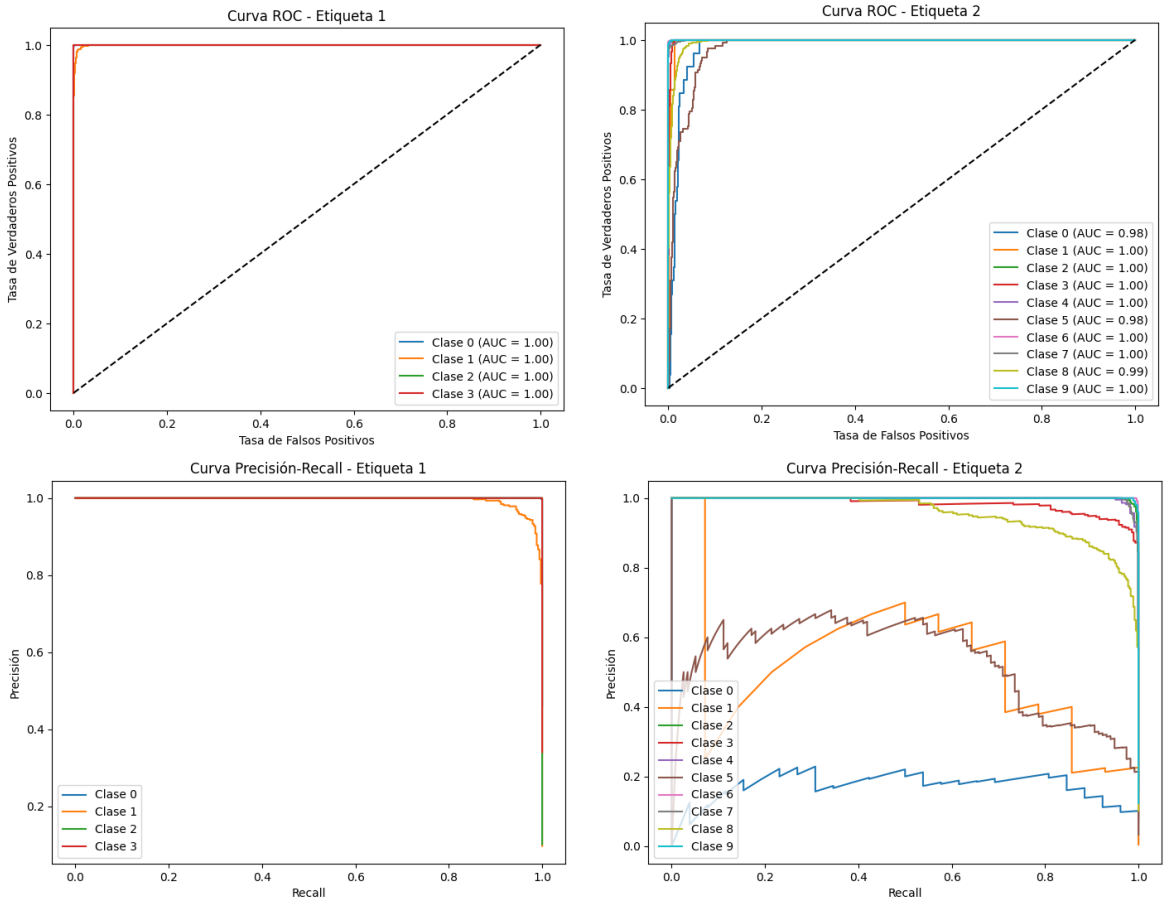

*Figura 11. **Evaluación del desempeño del modelo 4**: Los dos gráficos superiores muestran las Curvas ROC para la Etiqueta 1 y la Etiqueta 2, respectivamente, proporcionando una visualización del desempeño del modelo en términos de la tasa de verdaderos positivos frente a la tasa de falsos positivos. Los dos gráficos inferiores presentan las Curvas de Precisión/Sensibilidad para la Etiqueta 1 y la Etiqueta 2, respectivamente, mostrando la precisión del modelo en relación con la proporción de verdaderos positivos sobre el total de positivos reales y predichos.*

9 tienen un desempeño casi perfecto, mientras que las clases 0 y 1 presentan un rendimiento inferior con curvas más irregulares. La distribución de probabilidades (Figura 12) muestra una alta concentración cerca de 0, con picos menores alrededor de 0.01 y 0.015, sugiriendo que el modelo tiende a hacer predicciones con baja probabilidad para muchos casos.

Comparando ambos conjuntos de resultados, se observa que el modelo para la Etiqueta 1 muestra un rendimiento excepcionalmente alto en todas las métricas, cercano a la perfección. En contraste, el modelo para la Etiqueta 2, aunque muestra un buen desempeño general, presenta algunas áreas de mejora. La Etiqueta 2 muestra mayor variabilidad entre clases y algunas confusiones específicas que podrían requerir atención. Además, la tendencia del modelo de la Etiqueta 2 a producir probabilidades bajas para muchas predicciones sugiere una menor confianza general en comparación con el modelo de la Etiqueta 1. Estos resultados indican que el modelo de la Etiqueta 1 podría estar trabajando con datos más fácilmente separables, mientras que el modelo de la Etiqueta 2 enfrenta un problema de clasificación más desafiante.

Conclusiones

El presente proyecto ha desarrollado un método híbrido innovador que combina modelos de ensamblado (Extra Trees y Random Forest) con una red neuronal profunda (FFNN), maximizando las fortalezas de ambos enfoques. Esta arquitectura se ha diseñado específicamente para predecir secuencialmente dos etiquetas diferentes, lo que permite un aprendizaje más eficiente y robusto. El modelo aprovecha la salida de la primera etiqueta como entrada para la predicción de la segunda, facilitando el aprendizaje de las relaciones entre ambas tareas de clasificación.

El modelo demostró ser efectivo en la clasificación multiclase del Desorden Genético (4 clases) y en la clasificación más compleja de las Enfermedades (10 clases). La alta confiabilidad de las predicciones, junto con la capacidad de

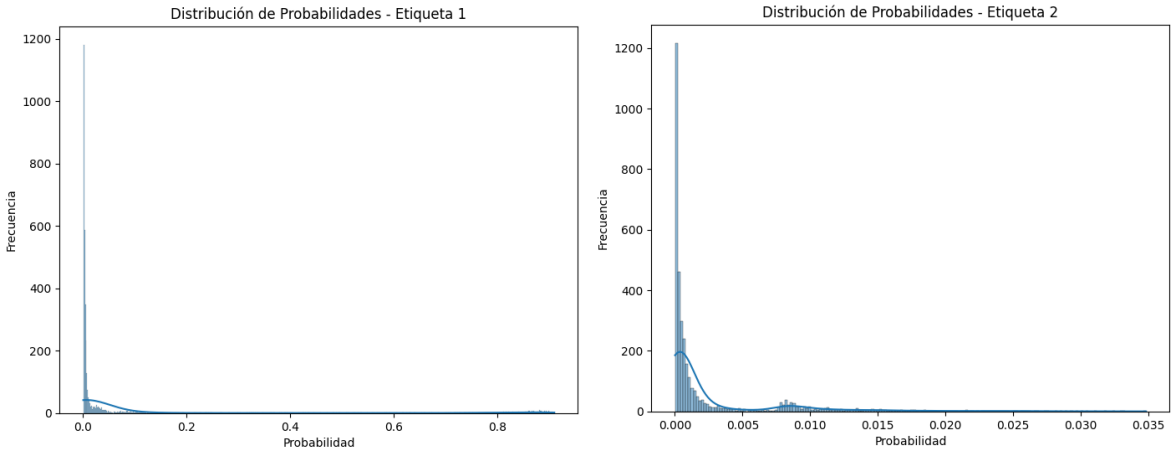

***Figura 12. Distribución de probabilidades para las etiquetas de clasificación**: Los gráficos muestran la distribución de probabilidades para las Etiquetas 1 y 2, respectivamente. Se muestra la frecuencia de las probabilidades predichas por el modelo de clasificación y permite evaluar la confianza en las predicciones de clasificación del modelo.*

generalización demostrada en los datos de prueba, sugiere que este enfoque tiene un gran potencial para aplicaciones prácticas en escenarios reales donde se requiera alta precisión en la clasificación de desórdenes genéticos.

Además, el enfoque utilizado para generar características iniciales mediante modelos de ensamblado, seguido de una red neuronal relativamente pequeña, ha demostrado un equilibrio óptimo entre rendimiento y eficiencia computacional. El uso de técnicas de regularización efectivas, como dropout y regularización L2, ha sido clave para prevenir el sobreajuste, como se evidencia en el rendimiento consistente del modelo en los datos de prueba.

La implementación del optimizador Adam con tasas de aprendizaje ajustables permitió una convergencia eficiente durante el entrenamiento, especialmente en los modelos más complejos. Además, la reducción de la complejidad del modelo en etapas posteriores del desarrollo no comprometió el rendimiento, lo que sugiere una arquitectura bien equilibrada.

En conclusión, el modelo desarrollado no solo cumple con el objetivo de predecir desórdenes genéticos y sus subclases en niños con alta precisión, sino que también muestra un excelente rendimiento general y potencial para utilizarse en aplicaciones prácticas en el campo del diagnóstico y tratamiento temprano de estas enfermedades. La robustez del modelo ante múltiples clases y su eficiencia computacional lo posicionan como una herramienta valiosa en el ámbito de la predicción médica.