

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

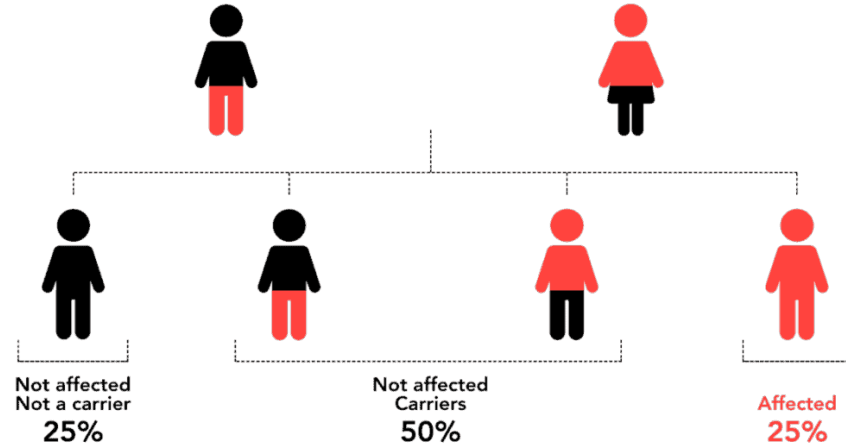


ANTECEDENTES

- ▶ El ADN contiene genes que son instrucciones para construir las proteínas que nuestro cuerpo necesita para funcionar.
- ▶ Los desórdenes genéticos son condiciones de salud que ocurren debido a alteraciones en el ADN de un individuo.
- ▶ Estos desórdenes pueden ser causados por una variedad de factores:
 - ▶ Algunos son heredados de los padres,
 - ▶ Otros pueden ser causados por cambios o mutaciones en los genes que ocurren durante la vida de una persona.

FATHER CARRIER

MOTHER CARRIER



CLASSIFICATION OF GENETICS DISORDERS

- Down syndrome
- Klinefelter syndrome
- Patau's syndrome (Trisomy 13)
- Edwards syndrome (Trisomy 18)
- Turner syndrome
- Cri du chat syndrome

CHROMOSOMAL
ABNORMALITIES

SINGLE
GENE
DISORDER

- A) X-LINKED
- Fragile X syndrome
- B) AUTOSOMAL RECESSIVE
- Congenital adrenal hyperplasia
- C) AUTOSOMAL DOMINANT
- Retinoblastoma

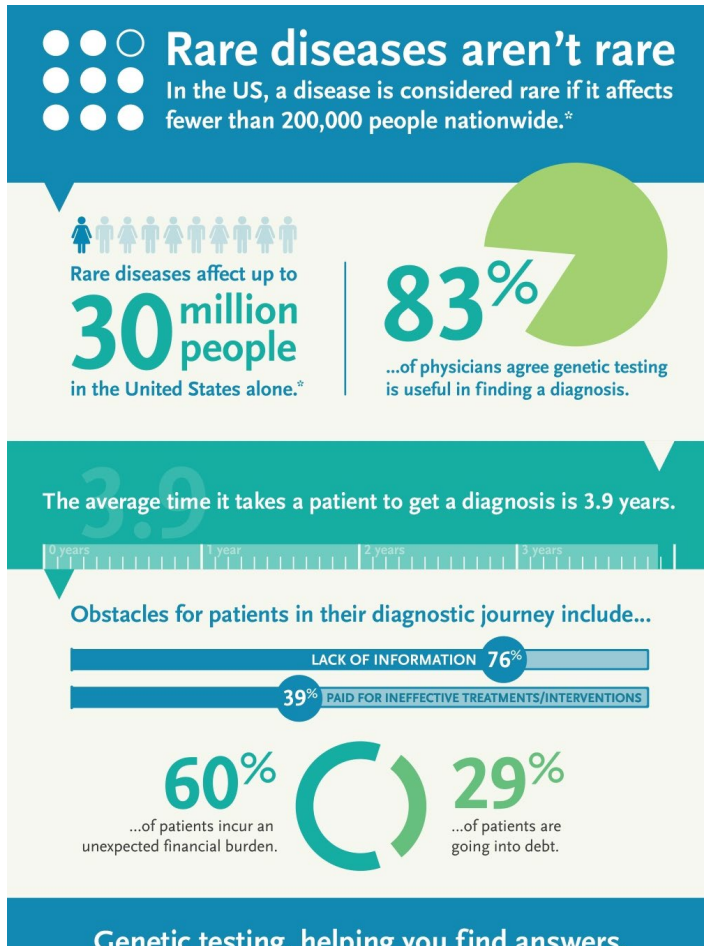
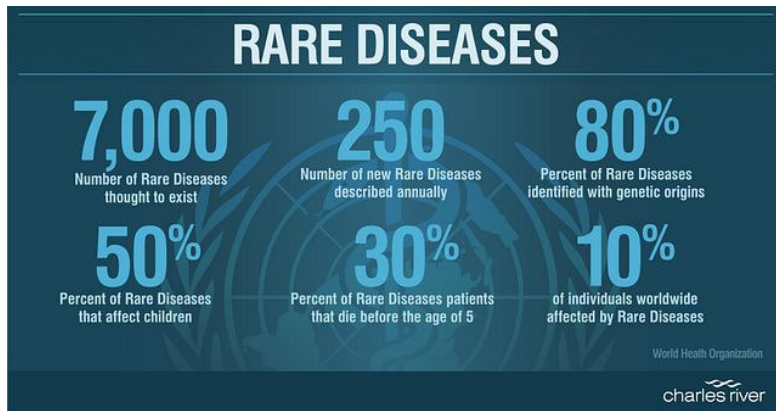
MULTIFACTORIAL
DISORDERS
CONGENITAL
MALFORMATION

MITOCHONDRIAL
DISEASES

- Cleft lip with or cleft palate
- Congenital heart defects
- Neural tube defect
- Mental Retardation

- Kaerns-Sayre syndrome
- Leber hereditary optic neuropathy
- Mitochondrial encephalopathy
- Myoclonic epilepsy

TIPOS DE DESÓRDENES GENÉTICOS



DESÓRDENES GENÉTICOS POCO FRECUENTES: ENFERMEDADES RARAS

- ▶ Las enfermedades genéticas poco frecuentes, raras, huérfanas o de baja prevalencia son enfermedades que se caracterizan por ocurrir a una baja frecuencia en la población general.
- ▶ Diferentes países tienen diferentes definiciones de prevalencia para enfermedades raras o poco frecuentes.
 - ▶ En América Latina y el Caribe, no todos los países cuentan con una definición de enfermedad rara o poco frecuente.
- ▶ La mayoría de las personas que viven con enfermedades raras y poco frecuentes no cuentan con el acceso a servicios médicos, sociales, de diagnóstico y tratamientos adecuados

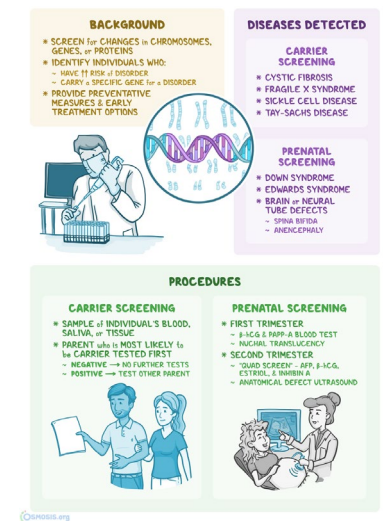
IMPACTO DE LOS DESÓRDENES GENÉTICOS



- ▶ El diagnóstico de una condición genética tiene un impacto profundo en la familia, pues puede experimentar sentimientos de culpa, depresión, miedo, entre otros.
 - ▶ Sin embargo, también puede tener un efecto de alivio al proporcionar un diagnóstico a una afección que antes era desconocida.
- ▶ El impacto de los desórdenes genéticos no se limita a la salud física, sino que puede tener efectos en diversas áreas de la vida.
 - ▶ Familiar, social, psicológicos, físicos, y económicos.

PRUEBAS GENÉTICAS

- ▶ La mayoría de las veces, los desórdenes genéticos se diagnostican a través de una prueba específica:
 - ▶ Examen de cromosomas o cariotipo
 - ▶ Prueba de sangre para ciertas enzimas que pueden ser anormales.
 - ▶ ADN (PCR o Secuenciación)



APRENDIZAJE AUTOMÁTICO PARA PREDICCIÓN DE DESÓRDENES GENÉTICOS



- ▶ Las técnicas de AA ofrecen herramientas poderosas para analizar y extraer conocimiento de estos conjuntos de datos complejos y heterogéneos.
 - ▶ Descubrimiento y desarrollo de medicamentos
 - ▶ Diagnóstico de enfermedades.
 - ▶ Análisis de imágenes médicas.
 - ▶ Análisis de Registros de Salud Electrónicos (EHR).
 - ▶ Análisis Genómico.

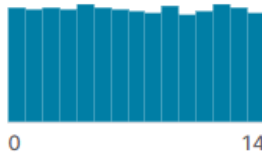



DEFINICIÓN DEL PROBLEMA

- ▶ 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.

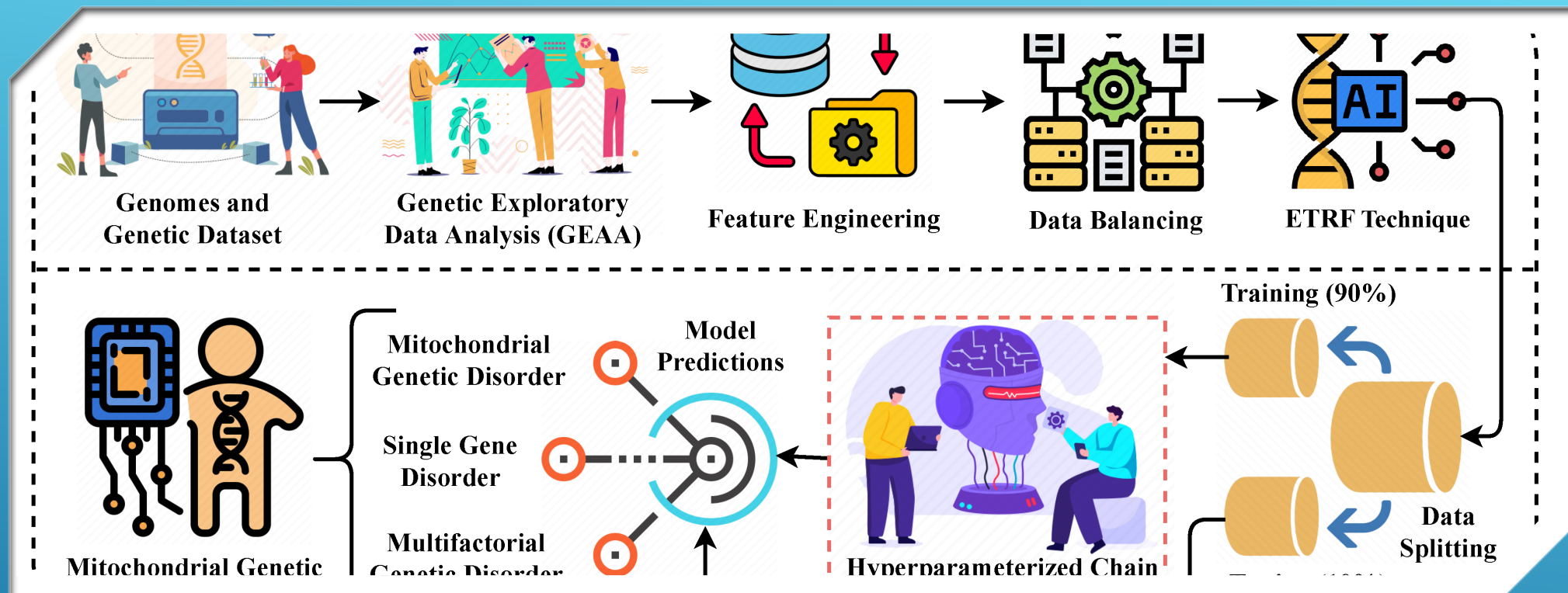
OBJETIVOS

- ▶ 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.

DESCRIPCIÓN DEL CONJUNTO DE DATOS

| ▲ Patient Id | # Patient Age | ✓ Genes in mother'... | ✓ Inherited from fa... | ✓ Maternal gene | | | | | | | | | | | | | | | | | | | | | | | | |
|------------------------|---|---|------------------------|-----------------|-----|-------|------|-----|--|------|------|-----|-------|-------|-----|--------|-----|----|--|------|-------|-----|-------|------|-----|--------|------|-----|
| 22083 unique values |  |  <table><tr><td>true</td><td>13.1k</td><td>60%</td></tr><tr><td>false</td><td>8940</td><td>40%</td></tr></table> | true | 13.1k | 60% | false | 8940 | 40% |  <table><tr><td>true</td><td>8644</td><td>39%</td></tr><tr><td>false</td><td>13.1k</td><td>59%</td></tr><tr><td>[null]</td><td>306</td><td>1%</td></tr></table> | true | 8644 | 39% | false | 13.1k | 59% | [null] | 306 | 1% |  <table><tr><td>true</td><td>10.6k</td><td>48%</td></tr><tr><td>false</td><td>8626</td><td>39%</td></tr><tr><td>[null]</td><td>2810</td><td>13%</td></tr></table> | true | 10.6k | 48% | false | 8626 | 39% | [null] | 2810 | 13% |
| true | 13.1k | 60% | | | | | | | | | | | | | | | | | | | | | | | | | | |
| false | 8940 | 40% | | | | | | | | | | | | | | | | | | | | | | | | | | |
| true | 8644 | 39% | | | | | | | | | | | | | | | | | | | | | | | | | | |
| false | 13.1k | 59% | | | | | | | | | | | | | | | | | | | | | | | | | | |
| [null] | 306 | 1% | | | | | | | | | | | | | | | | | | | | | | | | | | |
| true | 10.6k | 48% | | | | | | | | | | | | | | | | | | | | | | | | | | |
| false | 8626 | 39% | | | | | | | | | | | | | | | | | | | | | | | | | | |
| [null] | 2810 | 13% | | | | | | | | | | | | | | | | | | | | | | | | | | |

- ▶ El conjunto de datos contiene información médica sobre niños que tienen desórdenes genéticos. Consiste en archivos train.csv (22083 x 45), test.csv (9465 x 43) y sample_submission.csv (5 x 3).
- ▶ **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.
- ▶ **Características relevantes:** Incluye información genética, resultados de pruebas médicas, síntomas, edad de los padres, lugar de nacimiento, entre otros.
- ▶ **Objetivo de predicción:** Trastornos genéticos y sus subclases, ideal para un problema de clasificación supervisada.
- ▶ Permite trabajar en un problema con implicaciones reales en medicina genética.



Predicting Genetic Disorder and Types of Disorder Using Chain Classifier Approach

[Ali Raza](#), Conceptualization, Formal analysis, Writing – original draft,¹ [Furqan Rustam](#), Conceptualization, Data curation, Writing – original draft,² [Hafeez Ur Rehman Siddiqui](#), Methodology, Formal analysis, Supervision,¹ [Isabel de la Torre Diez](#), Resources, Project administration, Funding acquisition,^{3,*} [Begonia Garcia-Zapirain](#), Software, Investigation, Visualization,⁴ [Ernesto Lee](#), Software, Formal analysis, Data curation,⁵ and [Imran Ashraf](#), Validation, Writing – review & editing, Supervision^{6,*}

Andrey Sudarikov, Academic Editor

METODOLOGÍA GENERAL

The background features a dark blue geometric shape on the left and a lighter blue area on the right. Several large, semi-transparent question marks are scattered across the background. Three thin white lines cross diagonally from the bottom left towards the top right.

¿PREGUNTAS?

METODOLOGÍA GENERAL

Análisis exploratorio de datos

- Distribuciones de las características
- Visualización de datos
- Correlaciones
- Analisis de valores faltantes

Preprocesamiento de datos

- Limpieza de datos
- Manejo de valores faltantes
- Codificación de variables categóricas
- Normalización o estandarización de características

Ingeniería de características

- Creación de nuevas características a partir de las existentes
- Selección de características
- Sobremuestreo y/o Submuestreo

METODOLOGÍA GENERAL

Selección de modelos

- Entrenamiento de modelos. Probar dos modelos de aprendizaje supervisado: SVM o Redes Neuronales (CNN/DNN).
- Optimización

Evaluación de modelos

- Cálculo de métricas de rendimiento
- Comparación de modelos

| | | Frequently used algorithms for biomedical research | Example usage (data type) |
|-----------------------|--------------------------|--|--|
| Supervised learning | Machine learning | SVM | • Cancer vs healthy classification (gene expression) |
| | | KNN | • Multiclass tissue classification (gene expression) |
| | | Regression | • Genome-wide association analysis (SNP) |
| | Deep learning | Random forest | • Pathway-based classification (gene expression, SNP) |
| | | CNN | • Protein secondary structure prediction (amino acid sequence) |
| | | RNN | • Sequence similarity prediction (nucleotide sequence) |
| Unsupervised learning | Clustering | Hierarchical | • Protein family clustering (amino acid sequence) |
| | | K-means | • Clustering genes by chromosomes (gene expression) |
| | dimensionality reduction | PCA | • Classification of outliers (gene expression) |
| | | tSNE | • Data visualization (single cell RNA-sequencing) |
| | | NMF | • Clustering gene expression profiles (gene expression) |