Machine Learning: Predicting Genetic Disorders (Classification model)
Project proposal by **Salma Loukili & Nicolas Gutierrez**

Project Purpose:

The goal is to identify genetic disorders and their subclasses in children using a dataset filled with medical and familial information. This project aims to aid early detection and management of these disorders, which is crucial for improving the quality of life of affected children.

Dataset Overview:

Found in Kaggle:
https://www.kaggle.com/datasets/aibuzz/predict-the-genetic-disorders-datasetof-genomes/

The training data is 22,083 entries of children and 45 features each, including the known genetic disorders and subclasses.

Features included in Dataset:

- Patient Id.
- Genes in mother's side, Inherited from father, Maternal gene, Paternal gene
- Blood cell count (mcL): Important for diagnosing various genetic blood disorders.
- Parent's Ages: Can be correlated with increased risk for genetic mutations.
- Institute Name, Location of Institute
- Respiratory Rate, Heart Rate
- Tests 1-5
- Parental consent, Follow-up: Reflects the dataset's completeness and reliability.
- Birth asphyxia, Autopsy shows birth defect (if applicable).
- Folic acid details (peri-conceptional).
- History of serious maternal illness, radiation exposure, substance abuse.
- Assisted conception IVF/ART.
- History of anomalies in previous pregnancies, No. of previous abortions.
- White Blood Cell Count.
- Blood test result.
- Symptoms 1-5: Specific symptoms can be strong indicators of certain genetic disorders.
- Genetic Disorder, Disorder Subclass: The target variables for the model to predict.

Methodology:

1. Data Preparation: Clean and organize the dataset to ensure it's ready for analysis.
2. Exploratory Analysis: Understand the patterns and relationships in the data.
3. Feature selection: Use Random Forest for selecting the most relevant features for model training.
4. Model Training: Apply ML models on the training set, start with baseline Logistic regression and improve the performance further using decision trees, Random Forest and neural networks.
5. Model optimization: Perform tuning and testing until the optimal model is reached.

A portion of the training data will be used for testing. The model will use these features to learn from the training data, and make predictions of the testing data for feedback and optimization.

Interesting links:
Random Forest vs Decision Tree | Which Is Right for You? (analyticsvidhya.com)

https://www.kaggle.com/code/brsdincer/genomes-and-genetics-disorder-prediction-ii/notebook

https://www.kaggle.com/code/imsparsh/hackerearth-ml-of-genomes-and-genetics

https://www.mdpi.com/2073-4425/14/1/71