Problem Statement: Predicting the Genetic Disorders

by, Mitali Bansal

Date: Aug 13th, 2024

Capstone Project

UC Berkeley: PC-MLAI

Research Question

How can we accurately predict the type and subclass of genetic disorders in children based on their medical information and family history?

Expected Data Sources and Structure

Data Set URL: [Predict the Genetic Disorder]()

The dataset for this research will be sourced from the Kaggle with the title: "Predict the Genetic Disorder." The dataset contains medical information about children who have genetic disorders. It comprises 22,083 rows and includes the following data:

- Demographic Information
- Genetic Information
- Health Status and Medical Tests
- Consent and Follow-up
- Birth and Pregnancy Information

Expected Techniques

To achieve the expected results, the following techniques and methodologies will be employed:

Data Preprocessing:

- Handling missing values.
- Encoding categorical variables (e.g., converting Yes/No to 1/0).
- Normalizing and scaling numerical features.
- Removing duplicates and irrelevant columns.

Exploratory Data Analysis (EDA):

- Descriptive statistics to understand the distribution of data.
- Visualization techniques (scatter plots, heatmaps) to explore relationships between features.
- Correlation analysis to identify significant predictors.

Machine Learning Models:

- **Classification Algorithms**: Logistic Regression, Decision Trees, Random Forest, Gradient Boosting, and Support Vector Machines to build the predictive model.
- **Hyperparameter Tuning**: Grid search and cross-validation to optimize model performance.

Model Evaluation:

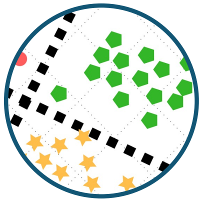
- Using performance metrics (accuracy, precision, recall, F1-score, AUC) to evaluate the models.
- Confusion matrix to visualize prediction results and identify misclassifications.

Feature Importance Analysis:

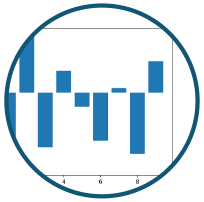
- Identifying and ranking features based on their contribution to the predictive model using techniques like Permutation Importance and SHAP (SHapley Additive exPlanations) values.

Expected Results

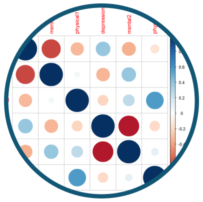
The goal of this research is to develop a predictive model that can accurately identify the type and subclass of genetic disorders in children. The expected outcomes include:



Classification Model: A robust machine learning model capable of predicting the genetic disorder and its subclass based on the given features.



Feature Importance: Identification of key features that significantly contribute to the prediction of genetic disorders.



Correlation Analysis: Understanding the relationship between various symptoms and the likelihood of specific genetic disorders.

Performance Metrics: Evaluation metrics such as accuracy, precision, recall, and F1-score to measure the effectiveness of the predictive model.

Why this question is important?

As per reports, because of the unsustainable increase in population and a lack of access to adequate health care, food, and shelter, the number of genetic disorder ailments have increased.

Hereditary illnesses are becoming more common due to a lack of understanding about the need for genetic testing. Often kids die because of these illnesses, thus genetic testing during pregnancy is critical.

This comprehensive approach aims to create a reliable predictive model that can assist healthcare professionals in early identification and management of genetic disorders, ultimately improving patient outcomes and advancing genetic research.

Thank you!