Capstone Two: "Of Genomes And Genetics" Project Documentation

Introduction

Objective: The purpose of this project is to apply machine learning techniques to predict genetic disorders based on clinical features. Genetic disorders often have complex traits that can be challenging to predict due to the variability in inheritance patterns and the influence of multiple genes. By developing predictive models, we aim to aid in the early diagnosis and understanding of these disorders.

Data Wrangling

Data from genetic studies, including patient demographics, genetic traits, and clinical signs, were loaded and cleaned. Missing values were handled appropriately, outliers were managed, and data consistency was ensured to prepare the dataset for analysis. For instance, missing age data was replaced with median values, and outlier ages were capped.

Exploratory Data Analysis (EDA)

Various visualizations were created to understand the relationships between different features and their impact on genetic disorders. Key insights include the distribution of age and its correlation with specific genetic disorders. For example, histograms and scatter plots highlighted the prevalence of certain disorders in specific age groups.

Pre-processing & Training Data Development

Data was pre-processed to ensure it could be effectively used for modeling. This included encoding categorical variables, normalizing numerical data, and splitting the dataset into training and testing subsets to evaluate the performance of the models.

Modeling

Several models were built and evaluated, including RandomForest and XGBoost. Model performance was measured using accuracy, precision, recall, and F1-score. The XGBoost model performed the best on the test data, with a detailed analysis of the confusion matrix and classification report helping to understand the model’s strengths and weaknesses.

Results

The final model achieved an accuracy of 55.4%, with the XGBoost model showing the best overall performance in terms of handling the imbalanced dataset. The model metrics comparison highlighted the need for further tuning and possibly gathering more data to improve the model's predictive capabilities.

Conclusion

This project demonstrated the potential of machine learning in predicting genetic disorders from clinical and genetic data. Despite the challenges of model sensitivity to imbalanced data, the insights gained from this analysis are valuable for further research and practical application in medical genetics.

Appendices

Code Snippets:

- Data cleaning and preprocessing
- Model training and evaluation metrics

Visualizations:

- Feature importance graph
- ROC curves for model comparisons

Model Metrics File

Random Forest:

- Accuracy: 53.8%
- Precision: 48%
- Recall: 54%
- F1-score: 45%

XGBoost:

- Accuracy: 55.4%
- Precision: 44%
- Recall: 55%
- F1-score: 42%