Executive Summary

This project aimed to analyze health-related data to predict disease categories based on various symptoms and indicators. Utilizing a Random Forest classifier, we developed a model capable of identifying the category of diseases with high accuracy. The methodology encompassed data preprocessing, feature engineering, model training, and hyperparameter tuning. Key findings revealed certain health indicators as critical predictors of disease categories. Recommendations include further exploration of feature importance to enhance preventive care and targeted treatments.

Introduction

The healthcare sector continuously seeks improved diagnostic tools to enhance patient outcomes. This project addresses the challenge of categorizing diseases based on symptoms and other health indicators, aiming to aid in early diagnosis and personalized treatment planning. By leveraging machine learning techniques, specifically a Random Forest classifier, we sought to predict disease categories accurately, contributing to the advancement of predictive healthcare analytics.

Data Overview

The dataset comprises health indicators and symptoms recorded from a diverse patient population, categorized into various diseases. It includes both numerical and categorical variables, such as age, blood pressure, cholesterol levels, and specific symptoms. Preprocessing steps involved handling missing values, encoding categorical variables, and addressing class imbalance through SMOTE technique and removing singleton classes to ensure model robustness.

Methodology

The project followed a structured data science workflow:

- Data Exploration: Initial analysis to understand the dataset's characteristics, distribution, and potential correlations.

- Preprocessing: Included cleaning data, encoding categorical variables, and addressing class imbalance.

- Feature Engineering: Grouped diseases into broader categories to reduce dimensionality and improve model interpretability.

- Model Training and Selection: Employed a Random Forest classifier for its efficacy in handling complex datasets with a mixture of feature types. Also attempted a Logistic regression first to establish a baseline.

- Evaluation: Used accuracy, confusion matrix, and cross-validation scores as primary metrics to assess model performance. Hyperparameter tuning was performed to optimize the model.

- Results: The model achieved an accuracy of 1.00 on the filtered dataset after removing singleton classes, indicating an excellent fit to the data. However, cross-validation confirmed the model's generalizability with consistent scores across folds. Feature importance analysis highlighted key indicators predictive of disease categories, offering insights into disease diagnostics and potential areas for medical research.

References:

https://www.kaggle.com/datasets/uom190346a/disease-symptoms-and-patient-profile-dataset