Doc Assist Project

Problem Statement:

The objective of this project is to develop an intelligent medical decision support system that analyzes patient data to assist doctors in making informed decisions about the best treatment options for individual patients. By leveraging machine learning and data analysis, the system will provide personalized treatment recommendations based on the patient's medical history, symptoms, lab results, and other relevant factors.

Overview

This code performs an analysis on a patients dataset and then builds a classification model using the Random Forest algorithm. The steps include importing necessary libraries, reading the dataset, performing exploratory data analysis (EDA), handling outliers, preprocessing the data, and training a Random Forest Classifier. Additionally, hyperparameter tuning is conducted using GridSearchCV, and the model is evaluated. Finally, a Logistic Regression model is trained for comparison.

Libraries Used:

- pandas: For data manipulation and analysis.
- numpy: For numerical operations.
- matplotlib and seaborn: For data visualisation.
- scipy.stats.zscore: For calculating z-scores for outlier detection.
- sklearn: For machine learning tasks, including data preprocessing, model training, and evaluation.
- joblib: For saving the trained model.

Dataset:

The dataset, loaded from "**dataset.xlsx**", contains haematology-related measurements such as *HAEMATOCRIT*, *HAEMOGLOBINS*, *ERYTHROCYTE*, *LEUCOCYTE*, *THROMBOCYTE*, *MCH*, *MCHC*, *MCV*, *AGE*, *SEX*, and *SOURCE*. SOURCE is the target variable indicating the source of the data. The dataset has 3309 entries with no missing values.

Exploratory Data Analysis (EDA):

EDA includes:

- Displaying the first few rows of the dataset.
- Checking data types and non-null counts.
- Visualisations such as pair plots, histograms, count plots for *SOURCE* and *SEX*, age distribution, age vs. *SOURCE*, and more.

Outlier Detection and Treatment:

Outliers are detected using z-scores, and a boxplot is used for visualisation. The outlier treatment function is implemented to remove outliers based on the IQR method.

Data Preprocessing:

The "**SEX**" column is mapped to numerical values (0 for 'M', 1 for 'F'). The dataset is then split into features (X) and target variable (y).

Model Training:

A Random Forest Classifier is trained on the pre-processed data. The model is evaluated using accuracy, a classification report, and a confusion matrix. Feature importances are also analysed and visualised.

Hyperparameter Tuning:

GridSearchCV is employed to find the best hyperparameters for the Random Forest model. The tuned model is then evaluated on the test set.

Logistic Regression Comparison:

A Logistic Regression model is trained and evaluated for comparison with the Random Forest model.

Results:

The Random Forest model achieves an accuracy of approximately 76.2%.

The Logistic Regression model achieves an accuracy of approximately 75.0%. Feature importance analysis suggests that *THROMBOCYTE*, *LEUCOCYTE*, and *HAEMATOCRIT* are crucial features for the classification task.

Model Persistence:

The Random Forest model is saved using joblib as 'model.joblib' for future use.

This detailed analysis and modelling process provides insights into the dataset and helps in building a predictive model for classifying the source of haematology data.

Discussion of future work:

In the context of the "Building Intelligent Decision Making Support System", there are several avenues for future work and improvements that can contribute to the advancement of medical research, diagnostics, and patient care:

1. Early Detection of Diseases:

Explore the potential for early detection of blood-related diseases. Identify patterns in the data that may indicate the onset of certain conditions, enabling proactive intervention and treatment.

2. Personalized Medicine:

Investigate the application of machine learning for personalized medicine. Develop models that consider individual patient characteristics to tailor treatment plans and interventions based on their unique profiles.

3. Improving Accuracy:

We can increase the accuracy by increasing the sample size of the dataset.