|  |
| --- |
| Anju Mohan  8-10-2024 |

|  |
| --- |
| **upGrad** |
| **Final Capstone Project** |
| **Doc Assist Project** |

**Doc Assist (Building Intelligent Medical Decision Support System)**

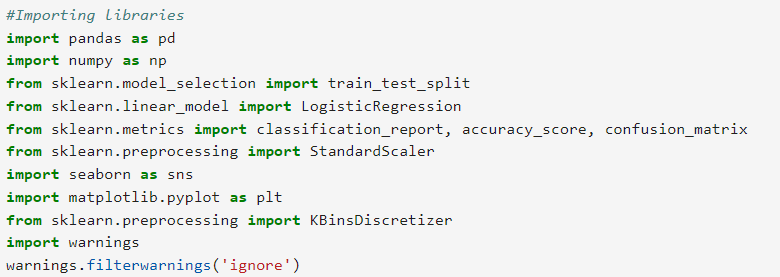
**Problem Statement**

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 conducts an analysis on a patient dataset and constructs a classification model utilizing the Random Forest algorithm. The process involves importing essential libraries, loading the dataset, performing exploratory data analysis (EDA), addressing outliers, preprocessing the data, and training a Random Forest Classifier. Additionally, hyperparameter tuning is executed using GridSearchCV, followed by model evaluation. Lastly, a Logistic Regression model is trained for comparison purposes.

**Libraries used**



**Data Set**

The loaded data set is dataset.xlsx. it contains 11 columns. 11 columns consist of Haematocrit, Haemoglobins, Erythrocyte, Leucocyte, Thrombocyte, MCH, MCHC, MCV, Age, Sex, Source. Source is the target variable. This data contains 3309 entries with 0 missing values.

**Exploratory Data Analysis**

EDA is a crucial step in data analysis that ensures you have a deep understanding of your dataset before moving on to more complex modeling tasks. EDA includes,

* **Understanding the data**

Display the top 5 lines of the data

Understanding the data types

Compute Summary statistics

Checking the non-null counts and null values

* **Visualisation of Data**

Visualise thedata using Histogram, bar chart, pair plot, count plots and heatmaps.

**Outlier Detection and Treatment**

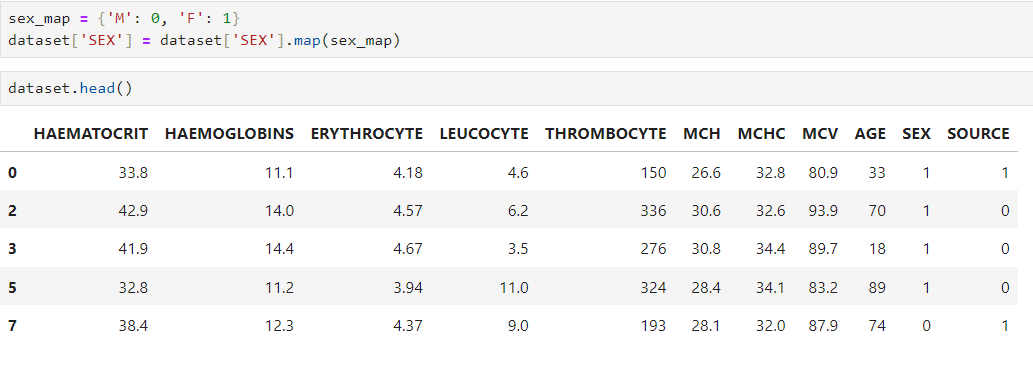
Outliers are data points that deviate significantly from the majority of the data. They can occur due to measurement errors, data entry errors, or genuine variability in the data. Identifying and handling outliers is crucial because they can skew the results of statistical analyses and machine learning models.

In this process, outliers are first detected using Z-scores and visualized with boxplots to understand their distribution. The outlier treatment is then carried out using the IQR method, which removes data points that fall outside a specified range. This combination of detection and treatment ensures that the data is clean, reliable, and ready for further analysis or modelling.

**Data Preprocessing**

It is a crucial step in the data analysis and machine learning pipeline. It involves transforming raw data into a clean and structured format that can be effectively used by models. Proper data preprocessing improves the accuracy and efficiency of the model by addressing issues such as missing data, inconsistent formatting, and irrelevant features.

Here 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).

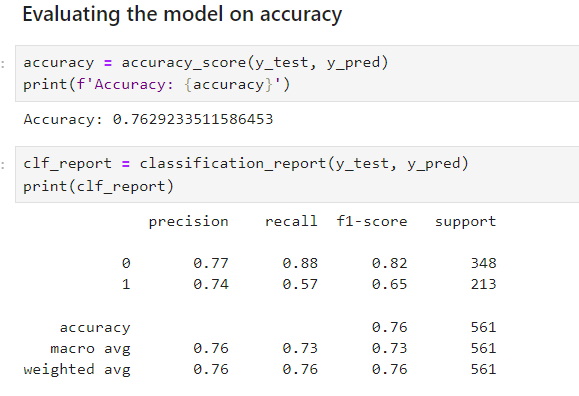


**Model Training**

Training a Random Forest Classifier model involves several key steps, including data preparation, model training, and evaluation. The Random Forest Classifier is an ensemble learning method that constructs multiple decision trees and merges their predictions to improve accuracy and control overfitting.

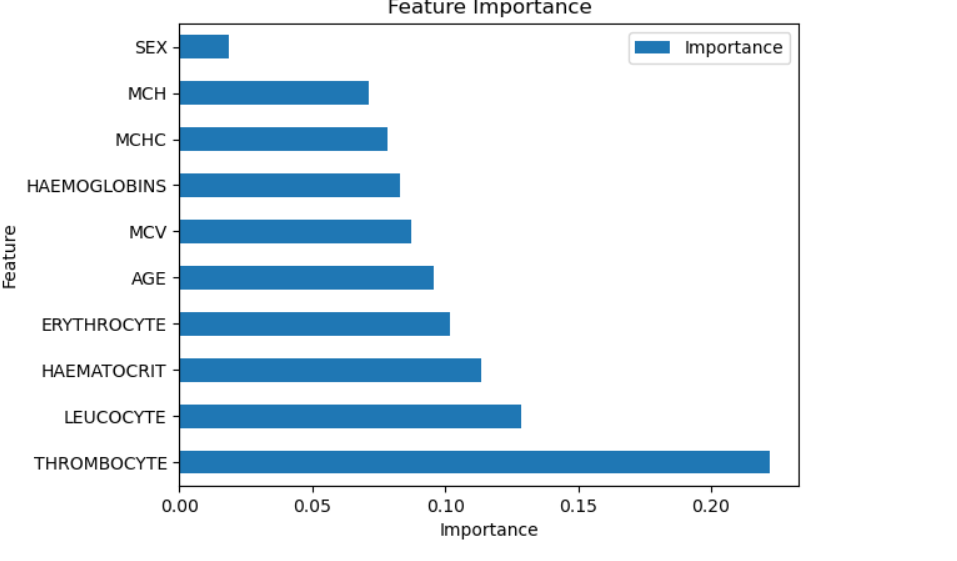
Steps include;

* Importing necessary libraries
* Loading the data
* Preprocess the data
* Instantiate the RandomForestClassifier with default parameters or specify any initial parameters.
* Fit the model on the training data
* Use the trained model to make predictions on the test set.
* Assess the model’s performance using metrics such as accuracy, precision, recall, F1-score, and confusion matrix.



**Feature Importance**

Feature importance is calculated based on the contribution of each feature to the reduction in impurity across all the trees in the forest. Feature importance helps in understanding which features contribute most to the predictions of the model. By analyzing feature importance, we can refine the feature set, improve model performance, and gain insights into the factors driving the model’s predictions.

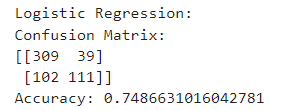


**Hyperparameter Tuning**

GridSearchCV is employed to find the best hyperparameters for the Random Forest model. This involves specifying a range of values for parameters such as n\_estimators, max\_depth, and min\_samples\_split. 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 76.29%.
* The Logistic Regression model achieves an accuracy of approximately 74.86%
* 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 for predicting with actual data.
* 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 scope of the "Building Intelligent Medical Decision-Making Support System," there are several promising directions for future research and enhancements that can significantly impact medical research, diagnostics, and patient care:

1. **Early Disease Detection:** Investigate the possibilities for the early identification of blood-related illnesses. Analyze data patterns that may signal the early stages of specific conditions, allowing for timely intervention and treatment.
2. **Personalized Healthcare:** Explore the use of machine learning to advance personalized medicine. Focus on developing models that take into account individual patient characteristics, enabling customized treatment plans and interventions tailored to their unique profiles.
3. **Enhancing Accuracy:** Consider improving accuracy by expanding the dataset. Increasing the sample size could lead to more precise and reliable predictions.
4. **Collaborative Research platform:** Create platforms that facilitate collaboration between researchers, clinicians, and data scientists. By sharing data, tools, and insights, these platforms can accelerate innovation and the development of new diagnostic and treatment approaches.
5. **Development of Real-Time Monitoring Systems:** Focus on creating systems that monitor patient data in real-time, allowing for continuous assessment and timely interventions. These systems could be particularly beneficial for managing chronic diseases or monitoring patients in critical care.
6. **AI-Driven Drug Discovery:** Investigate the application of AI and machine learning in drug discovery and development. By analyzing large datasets, AI can help identify potential drug candidates, predict drug responses, and accelerate the development of new treatments.
7. **Ethical and Transparent AI in Healthcare:** Address the ethical challenges of AI in healthcare by developing models that are transparent, explainable, and fair. Ensuring that AI-driven decisions can be understood and trusted by healthcare professionals and patients is crucial for widespread adoption.
8. **Predictive Analytics for Preventive Care:** Use predictive analytics to identify patients at risk of developing certain conditions and provide preventive care. By anticipating future health issues, healthcare providers can take proactive steps to mitigate risks and improve long-term health outcomes.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |