12/15/2023

****

**Prepared by**

Mitali Purohit

Tammy Xaypraseuth

Carlos Ramirez

**Supervised By**

Dr. Mohammad Pourhomayoun

Ashkan Aledavoud(Teaching Assistant)

Ly Jacky Nhiayi (Teaching Assistant)

**Identifying High-Risk Patients for Targeted Prevention and Intervention**

**CS – 4661 Introduction to Data Science**

Contents

[Project Overview 2](#_Toc153506807)

[Project Goals 2](#_Toc153506808)

[Team Members and Responsibilities 3](#_Toc153506809)

[Data Exploration and Preprocessing 3](#_Toc153506810)

[Dataset Description 3](#_Toc153506811)

[Data Preprocessing 4](#_Toc153506812)

[Feature Selection and Engineering 4](#_Toc153506813)

[Machine Learning Algorithms 4](#_Toc153506814)

[Decision Tree Algorithm 4](#_Toc153506815)

[Logistic Regression 4](#_Toc153506816)

[Random Forest Algorithm 5](#_Toc153506817)

[K-Nearest Neighbors (KNN) 5](#_Toc153506818)

[Linear Regression 5](#_Toc153506819)

[Cross-Validation Techniques 5](#_Toc153506820)

[Model Evaluation Metrics 5](#_Toc153506821)

[Feature Importance Analysis 5](#_Toc153506822)

[Challenges and Solutions 5](#_Toc153506823)

[Feature Work 6](#_Toc153506824)

[Conclusion 6](#_Toc153506825)

[References 7](#_Toc153506826)

## Project Overview

The "Identifying High-Risk Patients for Targeted Prevention and Intervention" project is a data-driven initiative within the healthcare domain. It seeks to leverage advanced data science and machine learning techniques to address a critical challenge facing the healthcare industry.

## Project Goals

**a. Predictive Patient Categorization:** Develop a robust machine learning model capable of accurately predicting and categorizing patients as either in-care or out-care. This predictive capability aims to enhance the efficiency of healthcare providers in managing patient treatments.

**b. Resource Allocation Optimization:** Optimize resource allocation within the hospital setting based on the predictive model's outcomes. By accurately categorizing patients, healthcare providers can streamline resource utilization, ensuring optimal patient care and resource efficiency.

**c. Integration with Hospital Automation System:** Integrate the predictive model seamlessly into the hospital automation system. This integration facilitates real-time decision-making, allowing healthcare professionals to make informed choices on patient care.

**d. Improved Patient Care Management:** Through predictive modeling, strive to improve overall patient care management by providing healthcare providers with valuable insights into patient conditions and treatment outcomes.

## Team Members and Responsibilities

**Mitali Purohit (Project Lead):** As the Project Lead, Mitali ensures the seamless execution of project objectives, orchestrating team efforts, and maintaining project timelines. She actively contributes to data preprocessing and model development.

**Tammy Xaypraseuth:** Tammy, the team's data exploration expert, plays a critical role in identifying key features and leading feature engineering activities. Her insights contribute significantly to the model's robustness.

**Carlos Ramirez:** Leading the modeling and algorithm development phase, Carlos brings his expertise to ensure the effective implementation and fine-tuning of machine learning models. His strategic approach enhances the practical applicability of our solutions.

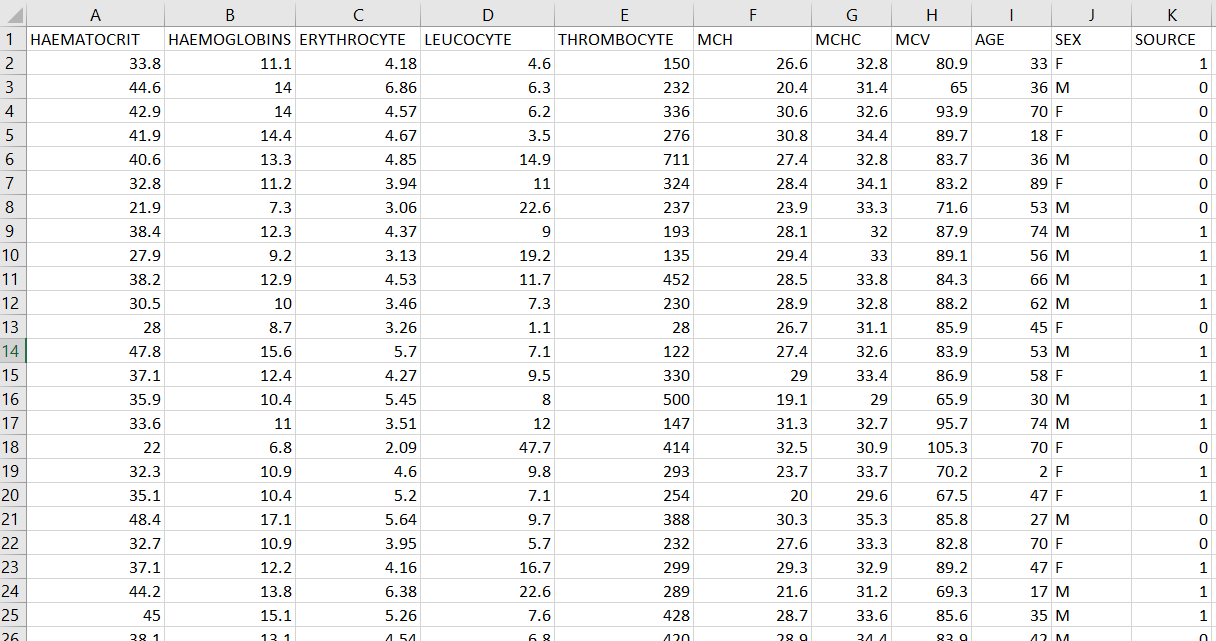
## Data Exploration and Preprocessing

### Dataset Description

**Dataset: Patient Treatment Classification**

**Source:** Private hospital in Indonesia

* **Content:** Patient conditions, lab test results, and other relevant factors
* **Availability:** [Kaggle Link](https://www.kaggle.com/datasets/manishkc06/patient-treatment-classification/data)



### Data Preprocessing

* Handling missing values and outliers.
* Encoding categorical variables.
* Scaling numerical features for optimal model training.

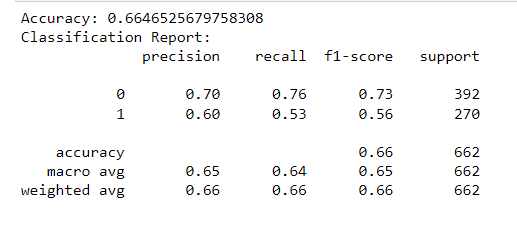
## Feature Selection and Engineering

In-depth exploration of features, selection of relevant variables, and creation of new features to enhance model performance.

## Machine Learning Algorithms

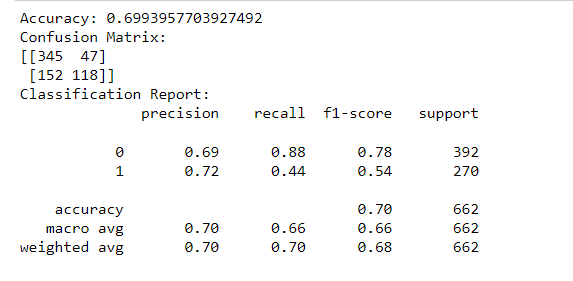
### Decision Tree Algorithm

* Tool: Scikit-learn library.
* Implementation: DecisionTreeClassifier.
* Results: ROC curves, AUC values, and feature importance analysis.



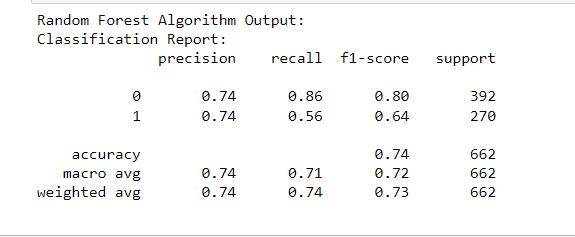
### Logistic Regression

* Tool: Scikit-learn library.
* Implementation: LogisticRegression.
* Results: ROC curves, AUC values, and interpretation of model coefficients.



### Random Forest Algorithm

* Tool: Scikit-learn library.
* Implementation: RandomForestClassifier.
* Results: Ensemble learning benefits, ROC curves, and AUC values.



### K-Nearest Neighbors (KNN)

* Tool: Scikit-learn library.
* Implementation: KNeighborsClassifier.
* Results: Pattern recognition and classification metrics.

Predictive Result Is:

[0.77777778]

### Linear Regression

* Tool: Scikit-learn library.
* Implementation: LinearRegression.
* Results: Regression analysis, model coefficients, and interpretation.
* Predictive Result Is :
* [0.49636383 0.25300368 0.29486361 ... 0.39888323 0.28209421 0.52865609]

## Cross-Validation Techniques

Exploration of cross-validation strategies to assess model performance and generalization.

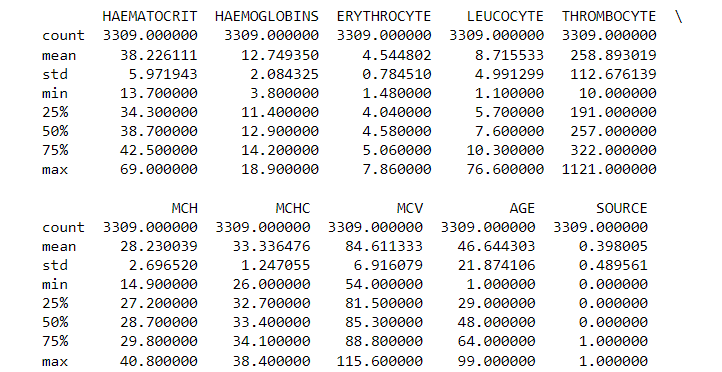
The machine learning model's performance was evaluated using cross-validation. The results are as follows:

* **Accuracy:** 0.75 (+/- 0.02)

This indicates the mean accuracy and the 95% confidence interval based on the cross-validation results.

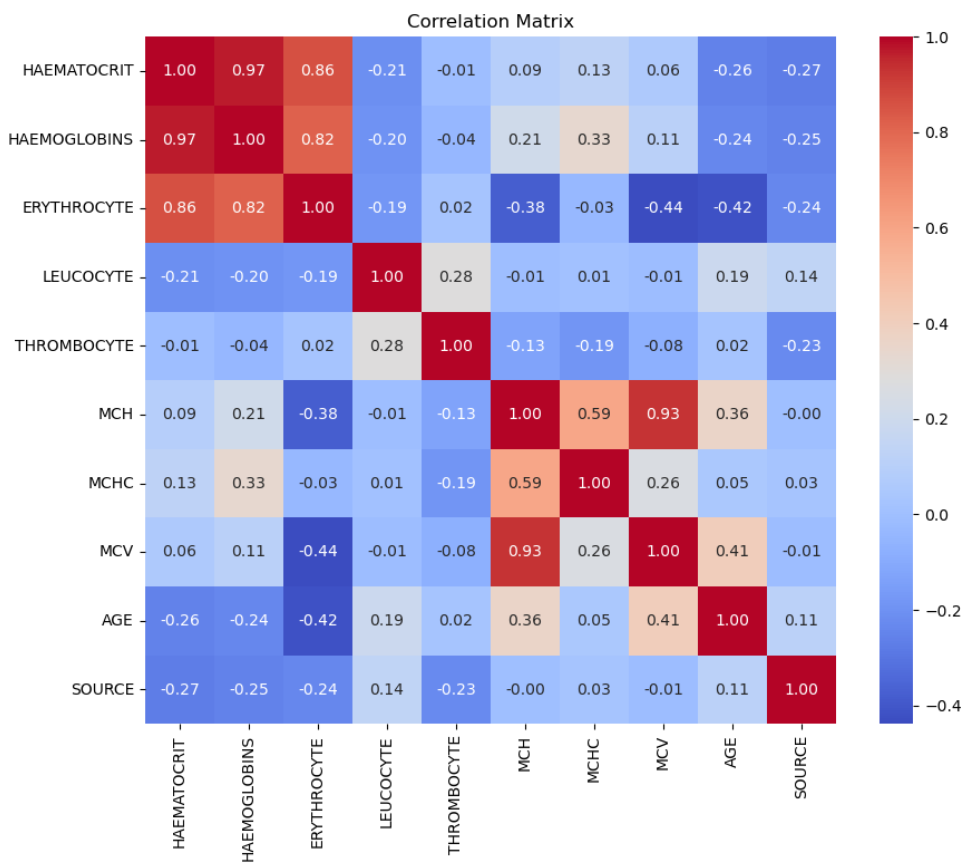
## Exploratory Data Analysis (EDA)

The provided code generates summary statistics for the numeric features in the dataset. This includes measures such as mean, standard deviation, minimum, maximum, and quartiles. These statistics offer a concise overview of the central tendencies and spread of the data.



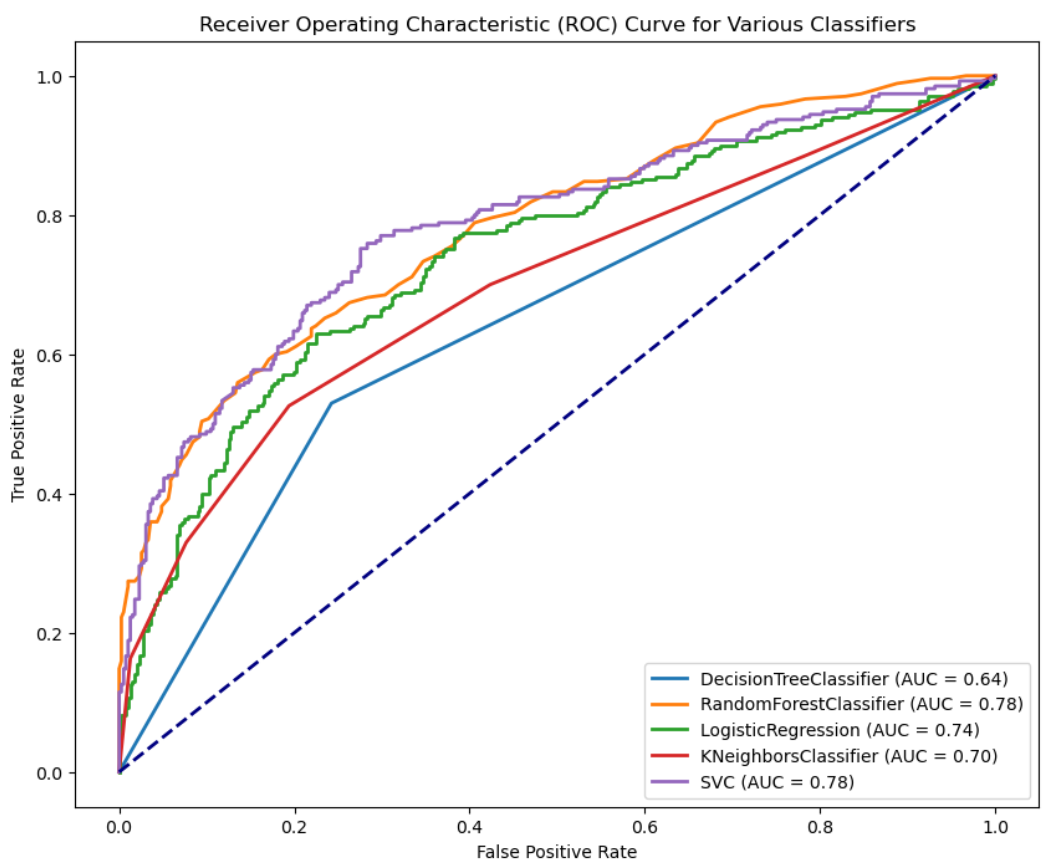
#### Correlation Matrix Heatmap

The Correlation Matrix visually represents the relationships between numeric variables in the dataset. Each cell displays the correlation coefficient, ranging from -1 to 1, indicating the strength and direction of the linear relationship. A value close to 1 suggests a positive correlation, while close to -1 implies a negative correlation. The heatmap aids in identifying patterns and dependencies among features, crucial for understanding potential predictive factors in the data.



## Model Evaluation Metrics

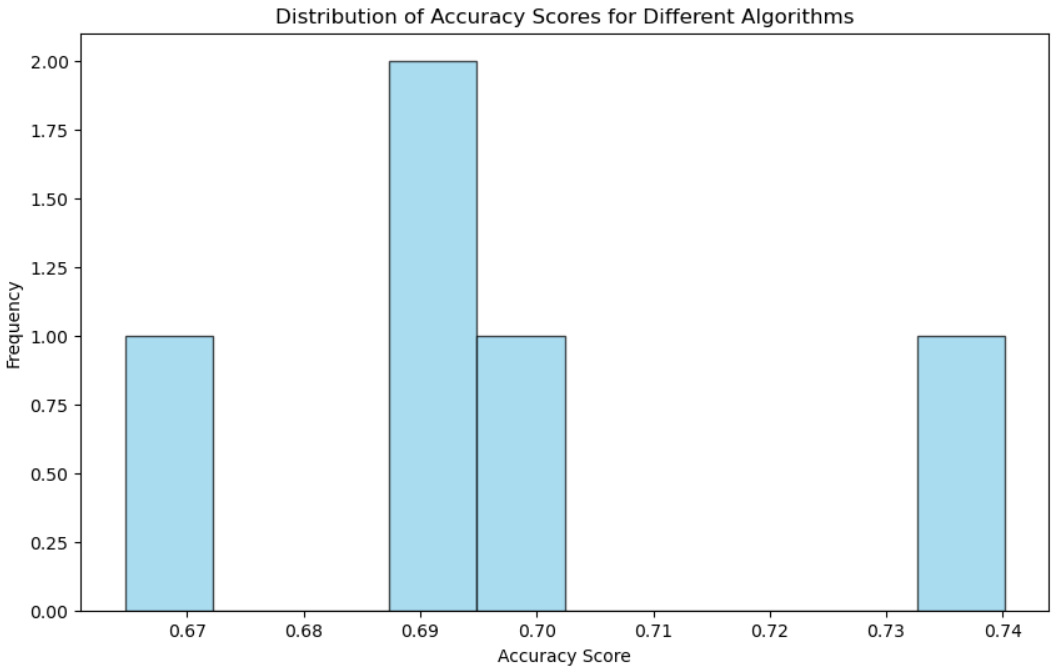
Comprehensive evaluation metrics, including ROC curves, AUC, accuracy, precision, recall, and F1-score.



## Comparative Accuracy Analysis

This section provides a visual overview of accuracy scores achieved by different machine learning algorithms. The histogram below depicts the distribution of accuracy scores from Decision Tree, Random Forest, Logistic Regression, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM). Each algorithm is color-coded, facilitating a quick comparison. Notable accuracy values include:

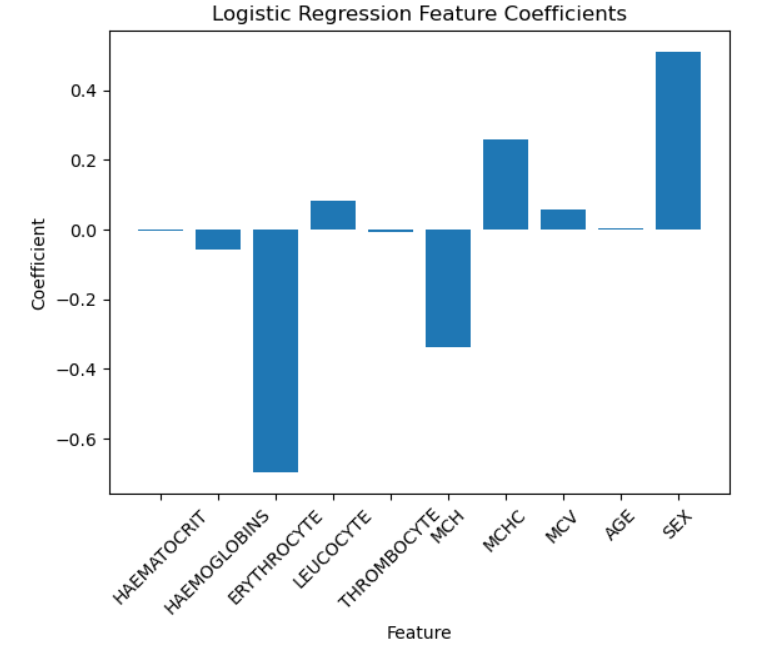
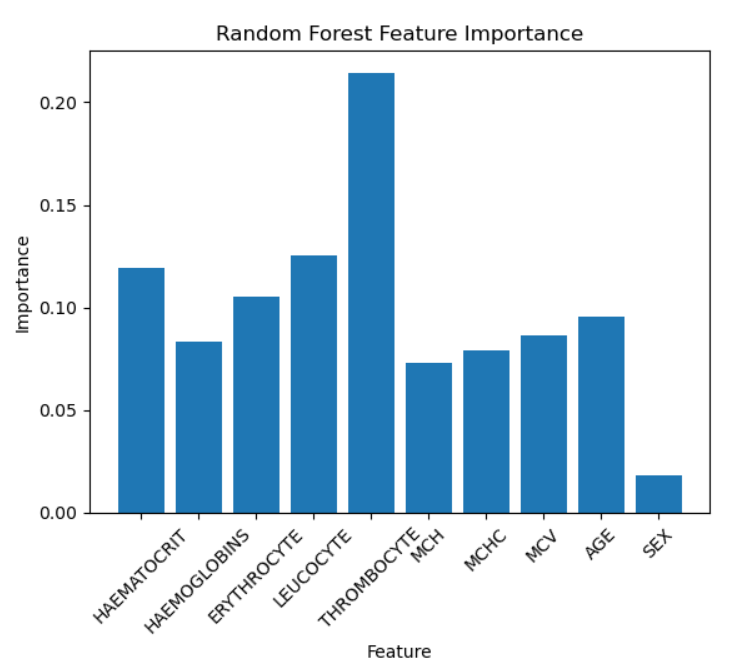
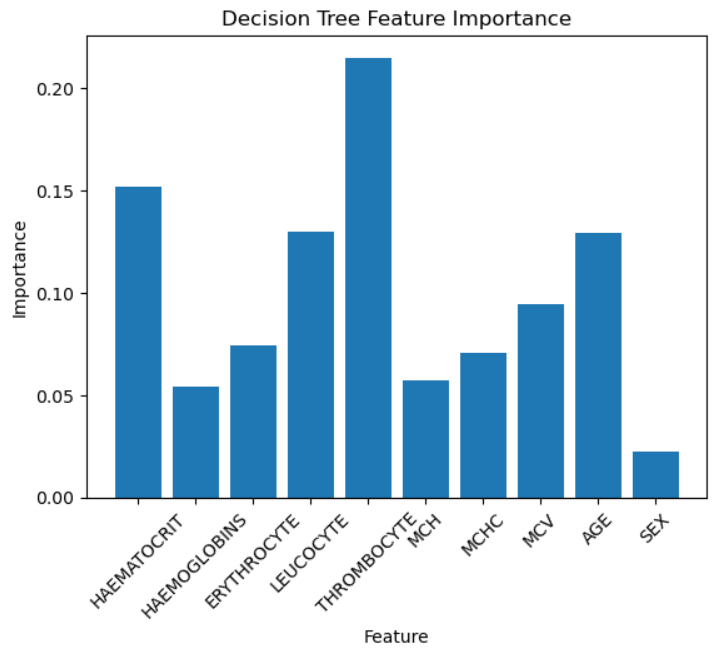
* Decision Tree: 0.6647
* Random Forest: 0.7402
* Logistic Regression: 0.6994
* K-Nearest Neighbors: 0.6918
* Support Vector Machine: 0.6918

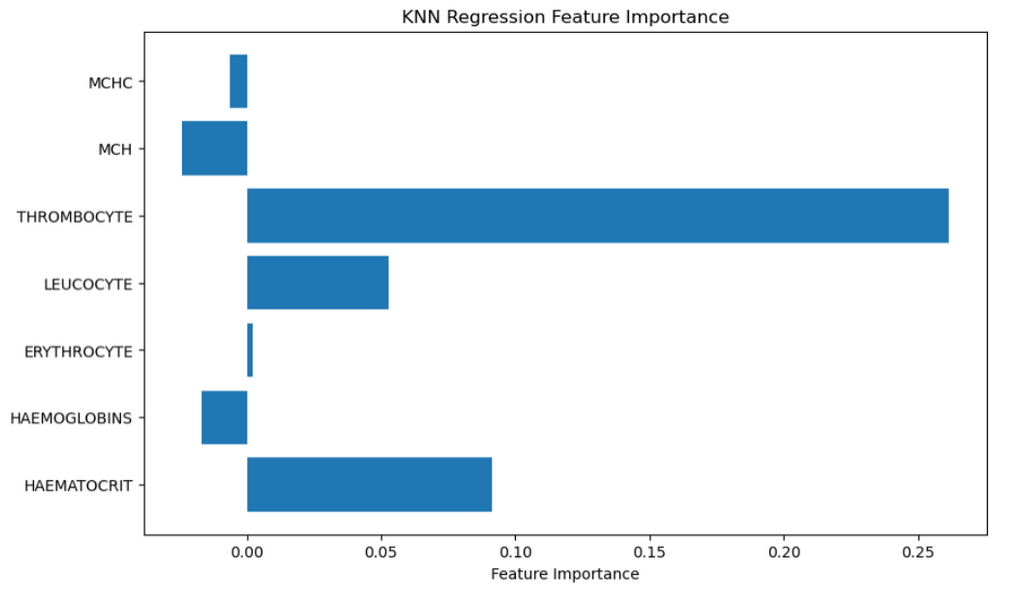
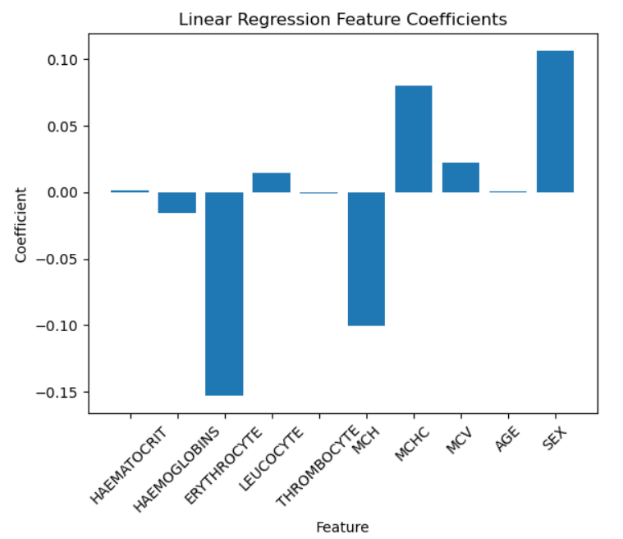


## Feature Importance Analysis

In-depth analysis of feature importance to interpret model decision-making processes.

Feature Importance Analysis provides insights into the significance of each feature in influencing the model's decision-making. By evaluating the contribution of individual features, this analysis identifies key predictors that strongly influence the model's output. Features with higher importance values play a more crucial role in the model's decisions, offering valuable information for understanding and interpreting the factors driving the predictive performance of the machine learning model.





## Challenges and Solutions

Addressing challenges encountered during the project, such as imbalanced data and hyperparameter optimization.

## Feature Work

The project has successfully achieved its primary goal of developing a predictive model for patient classification. Future work may involve refining the model further, exploring additional features, and considering real-world deployment challenges.

# Conclusion

The "Identifying High-Risk Patients for Targeted Prevention and Intervention" project represents an academic endeavor that holds substantial promise within the healthcare domain. By leveraging the power of data science and machine learning, the project aims to improve patient care, reduce healthcare costs, and enhance resource efficiency.

This project is available [at GitHub Repository](https://github.com/MitaliP001/Identifying-High-Risk-Patients-for-Targeted-Prevention-and-Intervention).

# References

1. Kaggle Datasets. (n.d.). "Patient Treatment Classification Dataset." Available online at: <https://www.kaggle.com/datasets/manishkc06/patient-treatment-classification>
2. M. Pourhomayoun, M. Shakibi, “[Predicting Mortality Risk in Patients with COVID-19 Using Artificial Intelligence to Help Medical Decision-Making](https://www.sciencedirect.com/science/article/pii/S2352648320300702?via%3Dihub#bib9),” the Journal of Elsevier Smart Health, 2020.
3. Python Software Foundation. (n.d.). "Python Programming Language." Available online at: <https://www.python.org/>
4. GitHub. (n.d.). "GitHub Documentation." Available online at: <https://docs.github.com/>