HEART DISEASE PREDICTION

Bhanu Pranaswi Sai P
S20210020309

Srinidhi M
S20210020323

Resumen—**ABSTRACT :**
This report explores a multimodal approach for Heart Disease Prediction, integrating Exploratory Data Analysis and machine learning techniques. Leveraging Linear regression, Logistic Regression, SVM, KNN, Random Forest, K means clustering, Feature Extraction, and Decision Tree models, alongside visualizations like training history plots and confusion matrices, we address the challenges of identifying misinformation. Data preprocessing includes handling outliers, Handling imbalanced data, PCA and feature scaling techniques, Box Cox Transformation, to enhance model robustness. Results demonstrate the effectiveness of the approach, providing insights for future research in the analysis of Heart Disease.

Index Terms—**Exploratory Data Analysis, Linear regression, Logistic Regression, SVM, KNN, Random Forest, K means clustering, Feature Extraction and Decision Tree models.**

I. INTRODUCTION

In the realm of healthcare, the prediction of heart disease stands as a critical endeavor with profound implications for public health and individual well-being. Amidst the rising prevalence of cardiovascular ailments, the ability to anticipate and mitigate the risk of heart disease has become increasingly vital. This imperative is underscored by the substantial burden that heart-related conditions impose on healthcare systems and communities worldwide. With heart disease remaining a leading cause of mortality globally, the development of our accurate predictive models holds the promise of early detection, timely intervention, and improved patient outcomes.

*I-0a. **The Significance of Heart Disease Prediction:**:* In recent years, the prevalence of heart disease has remained a pressing concern, with significant implications for public health and individual well-being. The ability to predict and preemptively address cardiac ailments has emerged as a pivotal endeavor in the healthcare landscape. By leveraging advanced analytics and predictive modeling techniques, we aim to enhance our capacity to identify individuals at risk of heart disease before symptoms manifest. This proactive approach not only enables timely interventions but also empowers individuals to make informed lifestyle choices aimed at reducing their risk factors. As we navigate the complexities of cardiovascular health, the development of accurate predictive models stands as a cornerstone in our efforts to combat heart disease and promote a healthier future for all.

*I-0b. **Motivation for the Study:**:* Studying heart disease prediction using real-world data is vital for public health, clinical decision-making, research, and economics. Early detection is crucial for improving outcomes, and predictive models assist in resource allocation and identifying high-risk individuals. They aid in risk assessment, treatment planning, and patient education. Research enhances understanding of risk factors and drives model development. Prevention leads to cost savings and productivity gains, making this study critical for healthcare and economic sustainability.

*I-0c. **Objectives and Methodology:**:* The study aims to develop a robust predictive framework for heart disease prediction through a multimodal approach integrating exploratory data analysis (EDA) and various machine learning techniques. Objectives include conducting comprehensive EDA to identify predictors, integrating diverse models like linear regression, logistic regression, SVM, KNN, random forest, and K-means clustering, and addressing challenges such as class imbalance and feature selection. Methodology involves data preprocessing, thorough EDA to understand dataset characteristics, model development and evaluation, advanced preprocessing techniques like outlier handling and feature scaling, and result interpretation for generating actionable insights. By systematically following these steps, the study endeavors to advance research in heart disease analysis and contribute to the broader understanding of cardiovascular health.

*I-0d. **Broader Implications:**:* Broader implications refer to the wider significance or consequences of a particular action, decision, or development beyond its immediate context or scope. In the context of a research project or study, broader implications often encompass the potential impacts, applications, or implications of the findings beyond the specific problem or domain being addressed. It involves considering how the outcomes of the research could affect society, industry, policy, or other relevant areas, and how they might contribute to broader knowledge, understanding, or advancement in the field. Identifying broader implications helps to contextualize the significance of the research and its potential relevance and importance in a larger context.

II. CONCEPTS USED:

Our approach encompasses a holistic methodology, from data preprocessing to model evaluation, to create a versatile and accurate Heart Disease Prediction system. By combining the strengths of various algorithms, we aim to contribute to the ongoing efforts to combat misinformation

1. Linear Regression: Used for predicting continuous numeric values, such as estimating heart disease risk based on various factors.

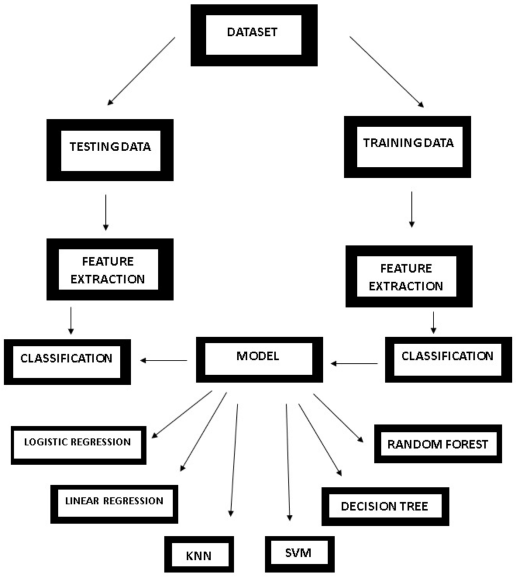


Figura 1. Block Diagram

2. Logistic Regression: Suitable for binary classification tasks, like predicting the presence or absence of heart disease based on patient data.

3. SVM (Support Vector Machine): Effective for both binary classification and regression tasks, SVMs can help classify heart disease risk or predict its severity.

4. KNN (K-Nearest Neighbors): Useful for classification based on similarity measures, KNN can be applied to predict heart disease based on similar patient profiles.

5. Random Forest: A versatile ensemble learning method, Random Forest can handle classification and regression tasks, making it suitable for predicting heart disease risk.

6. K-means Clustering: Primarily used for clustering and data segmentation, K-means can help identify patterns or groups within heart disease data.

7. Feature Extraction: Involves selecting or transforming relevant features from the data, which can enhance the performance of other models like Logistic Regression or SVM.

8. Decision Tree: Provides interpretable decision rules, useful for understanding factors contributing to heart disease risk and severity.

These concepts collectively contribute to creating a comprehensive and effective heart disease prediction.

III. CONCEPT APPLICATION AND OBTAINED RESULTS

EXPLORATORY DATA ANALYSIS: In our exploratory data analysis (EDA), we employed seaborn and matplotlib to visualize the dataset, focusing on univariate analysis to understand individual attribute distributions and identify outliers. Through these visualizations, we gained insights into the dataset's characteristics, discerning patterns and correlations between variables. Additionally, we utilized feature importance techniques to prioritize attributes crucial for heart disease prediction, ensuring the relevance of features used in model training.

Furthermore, our EDA encompassed feature scaling and transformation to enhance model robustness. By scaling attributes to a uniform range and applying transformations like PCA, we improved the data's interpretability and model performance. This thorough EDA process not only facilitated a deeper understanding of the dataset but also laid the groundwork for effective feature selection and model building in our heart disease prediction endeavor.

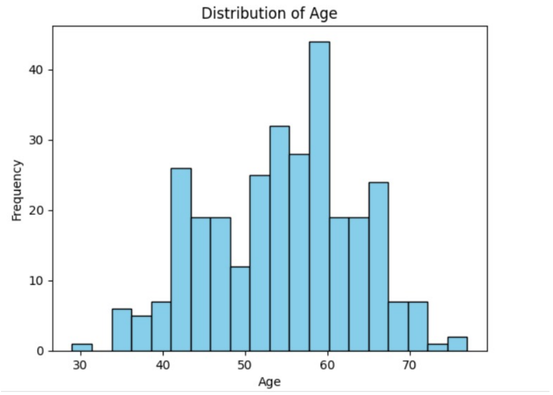


Figura 2. Frequency of Age - Histogram

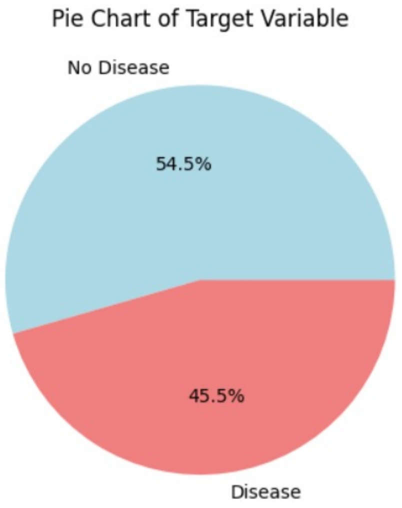


Figura 3. Pie chart

HANDLING THE OUTLIERS: In handling outliers, we employed visualization techniques such as box plots, kernel density estimation (KDE) plots, and violin plots to identify and understand the distribution of data points. These plots help visualize the spread and skewness of the data, allowing us to identify potential outliers based on their deviation from the central tendency. Ensuring the robustness and reliability of the dataset for further analysis.

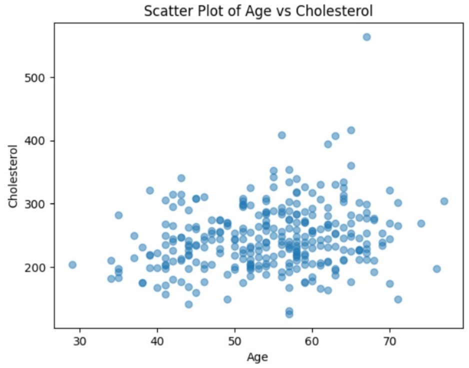


Figura 4. Scatter Plot

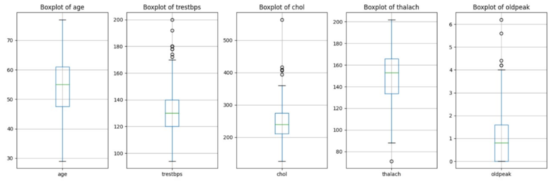


Figura 5. Box Plot

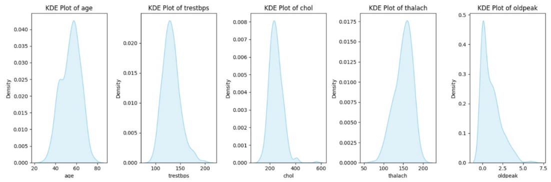


Figura 6. KDE Plot

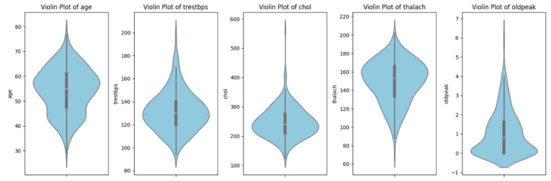


Figura 7. Violin Plot

FEATURE EXTRACTION:

Data Pre-processing: Before extracting features, the data is pre-processed to handle missing values, outliers, and noise.

Feature Importance: Feature importance analysis is conducted to identify the most relevant predictors of heart disease risk. This involves assessing the contribution of each feature to the prediction task using techniques like random forest feature importance or coefficient magnitudes in logistic regression.

Transforming Numerical Data: Numerical data, such as age, blood pressure, and cholesterol levels, are transformed using scaling techniques to bring them to a similar range and prevent features with larger magnitudes from dominating the learning process. Common scaling techniques include standardization (z-score normalization) and min-max scaling.

Dimensionality Reduction: Dimensionality reduction techniques like principal component analysis (PCA) are employed to reduce the number of features while preserving important information. This helps in reducing computational complexity and improving model efficiency without sacrificing predictive performance.

Feature Scaling: Feature scaling ensures that all features are on a similar scale, preventing bias towards features with larger magnitudes during model training. Scaling techniques are applied to numerical features to standardize their ranges and improve model convergence and performance.

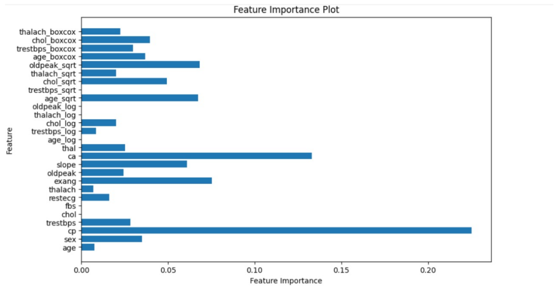


Figura 8. Feature Importance plot

CORRELATION RESULTS :

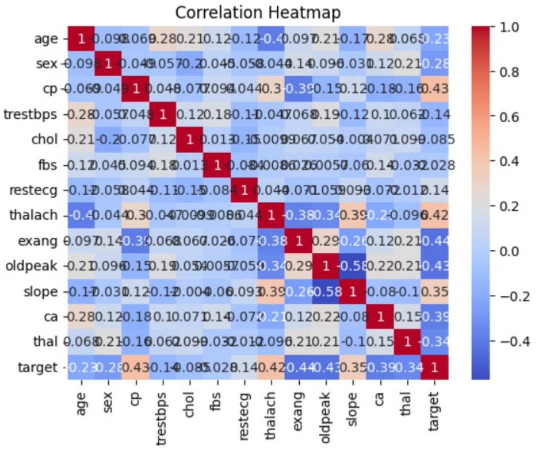


Figura 9. Correlation

In heart disease prediction, a correlation matrix provides crucial insights into the relationships among different features and their connection to the likelihood of heart disease occurrence. The matrix reveals the degree of correlation between variables, helping identify patterns and dependencies within the dataset related to heart health. By examining feature-target correlations, a clearer understanding of which characteristics are strongly associated with the risk of heart disease is gained. This information guides feature selection for machine learning models, focusing on the most relevant indicators of heart disease risk. Additionally, correlations expose potential issues like multi-collinearity, indicating when features convey similar information. Effectively interpreting the correlation matrix aids in building more accurate and robust models for heart disease prediction, contributing to better healthcare outcomes.

CONFUSION MATRICES:

Confusion matrix is a fundamental tool in evaluating the performance of classification models, providing insights into model accuracy, precision, recall, and other metrics. It consists of four quadrants: true positives, true negatives, false positives, and false negatives. True positives represent correctly predicted positive instances, while true negatives represent correctly predicted negative instances. False positives are instances incorrectly classified as positive, and false negatives are instances incorrectly classified as negative. Analyzing the confusion matrix allows us to calculate performance metrics such as accuracy, precision, recall, and F1 score, providing a comprehensive assessment of the model's predictive capabilities and potential areas for improvement.

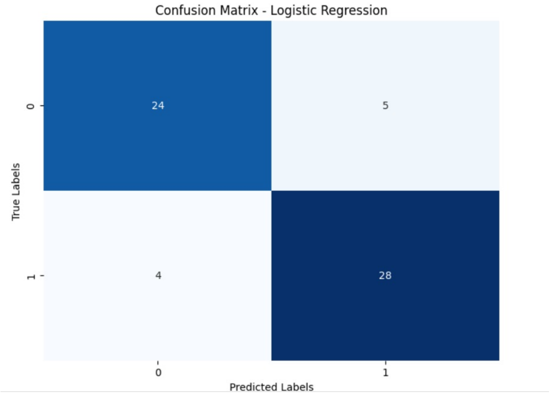


Figura 10. CONFUSION MATRIX(LOGISTIC REGRESSION)

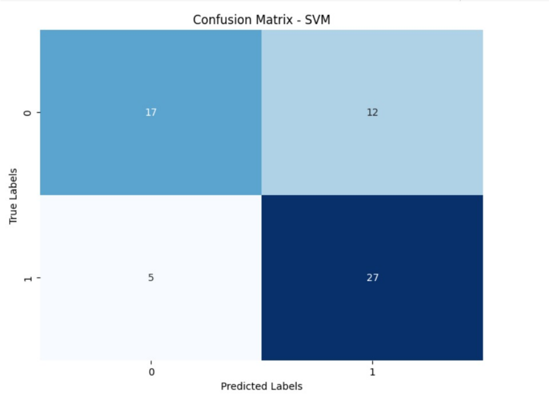


Figura 11. CONFUSION MATRIX(SVM)

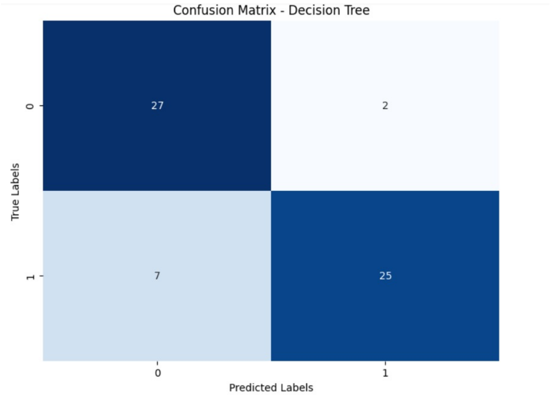


Figura 12. CONFUSION MATRIX(DECISION TREE)

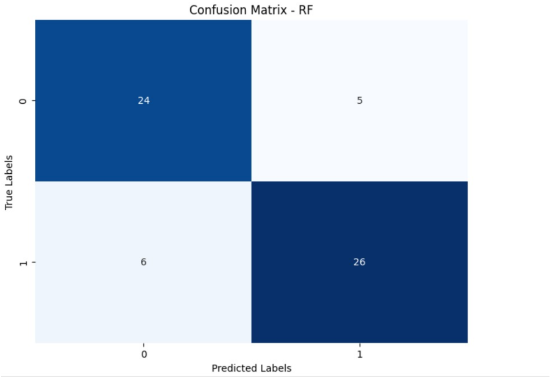


Figura 13. CONFUSION MATRIX(RANDOM FOREST)

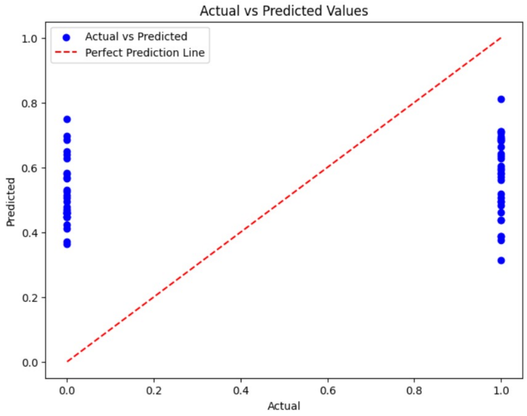


Figura 14. Linear Regression

- **EDA (Exploratory Data Analysis):** EDA involves analyzing and visualizing the dataset to understand its characteristics and uncover patterns or trends. In heart disease prediction, EDA helps identify relationships between variables, detect outliers, and inform feature selection and model building decisions.
- **Feature Extraction:** Feature extraction involves selecting or transforming relevant features from the data to enhance the performance of other models like Logistic Regression or SVM. In heart disease prediction, feature extraction techniques can help identify the most informa-

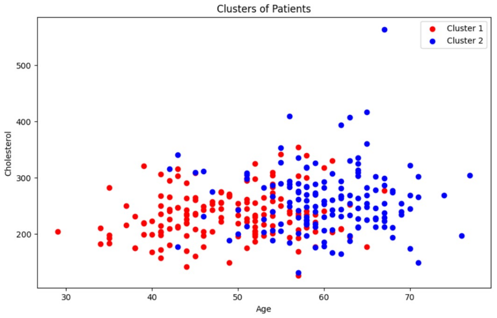


Figura 15. K-means

tive predictors and improve model accuracy.

*III-0a. **Application:*** Feature extraction techniques in heart disease prediction involve extracting informative features from patient data, such as age, blood pressure, and cholesterol levels, to enhance prediction accuracy and interpretability.

- **K-means Clustering:** K-means clustering is primarily used for data segmentation and identifying patterns within heart disease data. It can help group patients into clusters based on similar characteristics, aiding in understanding the underlying structure of the data and identifying potential risk factors.
- **Linear Regression:** Linear regression is used to predict continuous numeric values, such as estimating heart disease risk based on various factors like age, blood pressure, and cholesterol levels. It establishes a linear relationship between the predictor variables and the target variable.

*III-0b. **Application:*** Linear regression is commonly used in predicting outcomes, such as stock prices in finance, sales volumes in business, or disease prevalence in medical research.

- **SVM (Support Vector Machine):** SVM is effective for both binary classification and regression tasks in heart disease prediction. It finds the optimal hyperplane that best separates classes in the feature space, allowing it to classify heart disease risk or predict its severity accurately.

*III-0c. **Application:*** SVM is effective for text classification tasks, as it can handle high-dimensional data and is particularly useful when dealing with non-linear relationships between features.

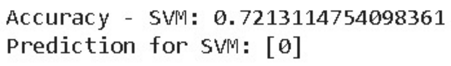

```
Accuracy - SVM: 0.7213114754098361
Prediction for SVM: [0]
```

Figura 16. ***Support Vector Machine***

*III-0f. **Application:*** Decision trees are interpretable and useful for understanding the decision-making process. They can be employed for classification tasks and feature importance analysis.

- **KNN (K-Nearest Neighbors):** KNN is useful for classification based on similarity measures. In heart disease prediction, KNN can predict the likelihood of heart disease based on the similarity of patient profiles to those with known outcomes, making it a valuable predictive model.

*III-0d. **Application:*** KNN is suitable for classification tasks, providing a non-parametric approach to making predictions based on the similarity of instances.

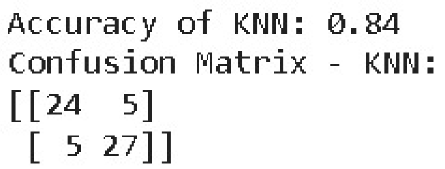

```
Accuracy of KNN: 0.84
Confusion Matrix - KNN:
[[24  5]
 [ 5 27]]
```

Figura 17. ***K-Nearest Neighbor***

- **Random Forest:** Random Forest is a versatile ensemble learning method capable of handling classification and regression tasks. In heart disease prediction, it can effectively predict heart disease risk by aggregating predictions from multiple decision trees, providing robust and accurate results.

*III-0e. **Application:*** Random Forest is robust and effective for classification tasks. It helps to overcome overfitting and enhances the model's generalization performance.

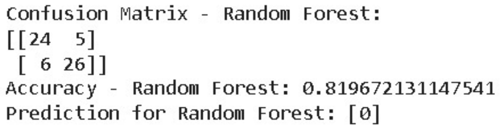

```
Confusion Matrix - Random Forest:
[[24  5]
 [ 6 26]]
Accuracy - Random Forest: 0.819672131147541
Prediction for Random Forest: [0]
```

Figura 18. ***Random Forest***

- **Decision Tree:** Decision trees provide interpretable decision rules, making them useful for understanding factors contributing to heart disease risk and severity. They partition the feature space based on the most informative features, allowing for easy interpretation of the model's predictions.

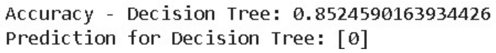

```
Accuracy - Decision Tree: 0.8524590163934426
Prediction for Decision Tree: [0]
```

Figura 19. ***Decision Tree***

- **Logistic Regression:** Logistic regression is suitable for binary classification tasks, such as predicting the presence or absence of heart disease based on patient data. It estimates the probability that a given instance belongs to a particular class, making it useful for disease prediction.

*III-0g. **Application:*** Logistic regression is applied in heart disease prediction to classify patients as either having or not having the disease based on their clinical characteristics and risk factors.

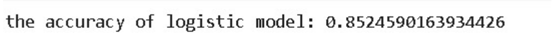

```
the accuracy of logistic model: 0.8524590163934426
```

Figura 20. ***Logistic Regression***

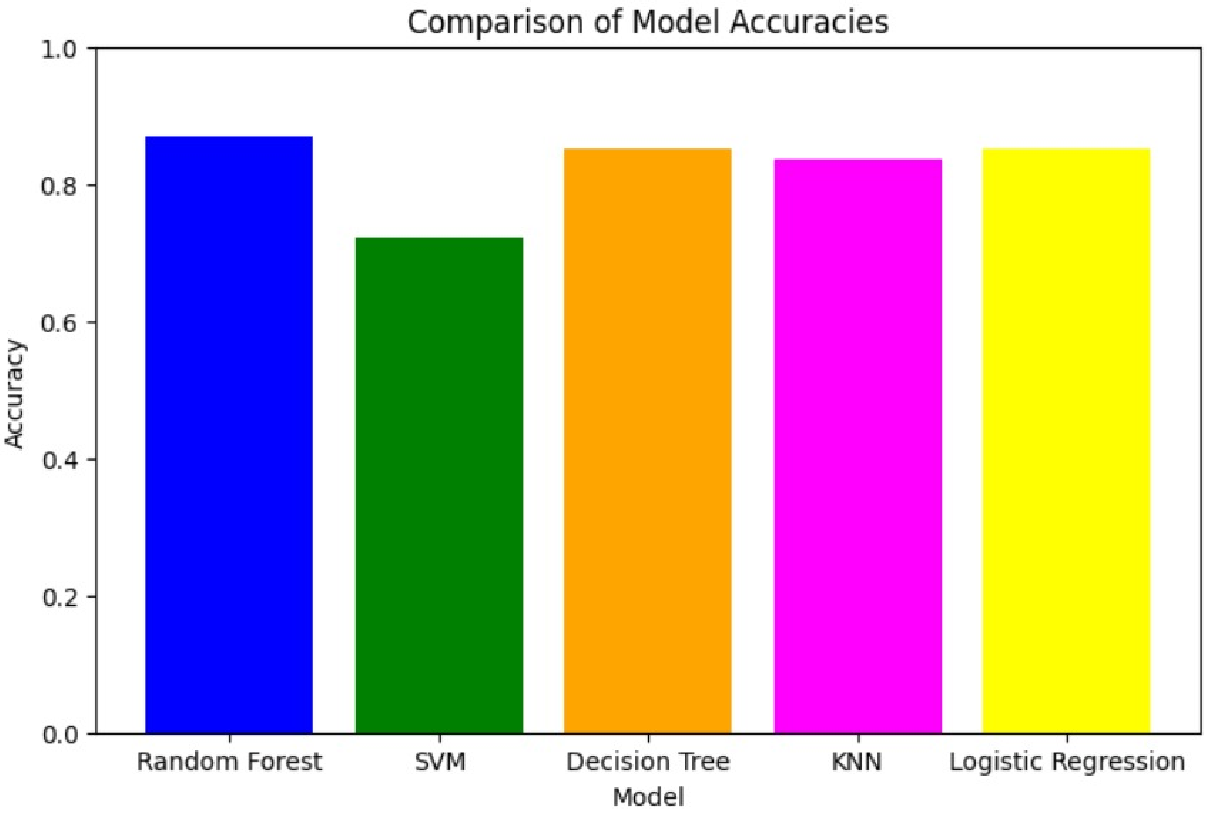
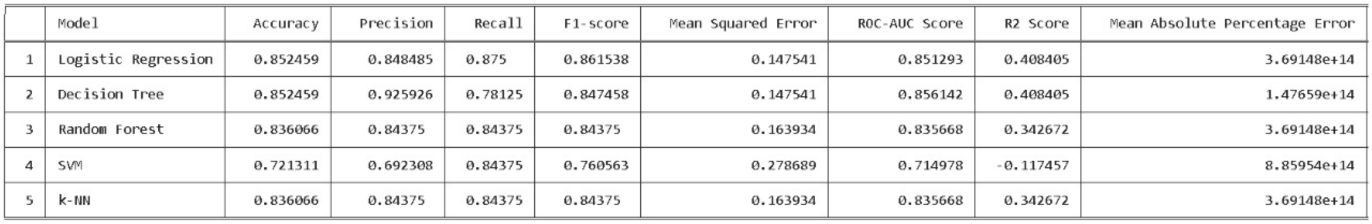


Figura 21. Comparing all the Models

MODEL RESULTS:



	Model	Accuracy	Precision	Recall	F1-score	Mean Squared Error	ROC-AUC Score	R2 Score	Mean Absolute Percentage Error
1	Logistic Regression	0.852459	0.848485	0.875	0.861538	0.147541	0.851293	0.408405	3.69148e+14
2	Decision Tree	0.852459	0.925926	0.78125	0.847458	0.147541	0.856142	0.408405	1.47659e+14
3	Random Forest	0.836066	0.84375	0.84375	0.84375	0.163934	0.835668	0.342672	3.69148e+14
4	SVM	0.721311	0.692308	0.84375	0.760563	0.278689	0.714978	-0.117457	8.85954e+14
5	k-NN	0.836066	0.84375	0.84375	0.84375	0.163934	0.835668	0.342672	3.69148e+14

Figura 22. COMPARISION BETWEEN ALL THE MODELS

IV. Conclusion

In our heart disease prediction project, we applied a range of sophisticated techniques to achieve accurate results. This included utilizing Support Vector Machines (SVM), Decision Trees (DT), Random Forest (RF), Logistic Regression (LogReg), and k-Nearest Neighbors (KNN) models. We conducted thorough Exploratory Data Analysis (EDA) to understand the dataset better. To address data distribution and scaling issues, we employed Box-Cox transformation and MinMax scaling techniques. Additionally, we tackled class imbalance using SMOTE (Synthetic Minority Over-sampling Technique) and gained insights through k-means clustering for data segmentation and analysis. These approaches collectively contributed to a comprehensive and robust heart disease prediction Analysis.