ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24SEP1684

Optimizing Heart Disease Diagnosis: Feature Selection Techniques for Enhanced Machine Learning Model Performance

Ravinder Kaur[1]; Sonia Rani[2*]; Chitra Desai[3]; Sagar Jambhorkar[4]
[1,2,3,4]Department of Computer Science, National Defence Academy, Pune, India

Corresponding Author: Sonia Rani[2*]

Abstract:- Heart disease is a growing global concern, affecting people across various age groups and genders. Detecting heart failure early is crucial, and ongoing research leverages advancements in healthcare technology, machine learning, imaging techniques, and data science to analyze vast datasets for this purpose. However, not all data attributes contribute equally to diagnosing heart disease, and the inclusion of irrelevant features can increase resource demands and potentially lead to inaccurate predictions with fatal consequences. This study focuses on feature extraction and reduction techniques to identify the most critical attributes for heart disease diagnosis, balancing resource efficiency with diagnostic accuracy. Using a dataset from the UCI repository, which includes both continuous and categorical features, we standardize the data and split it into training and testing sets in an 80:20 ratio. We then apply feature selection techniques to machine learning models such as K-nearest neighbor, decision tree classifier, SVM, logistic regression, and random forest. The models' predictive performance is evaluated using confusion matrices and ROC curves, demonstrating the impact of feature selection on diagnostic accuracy.

I. INTRODUCTION

Heart disease remains a leading cause of mortality worldwide, affecting millions of individuals across diverse demographics. Early detection and accurate diagnosis of heart failure are critical for improving patient outcomes, reducing healthcare costs, and guiding treatment decisions. The increasing availability of large-scale healthcare data and advancements in machine learning (ML) and data science provide unprecedented opportunities to enhance heart disease diagnosis through data-driven insights.

The challenge in heart disease diagnosis lies not only in the accurate prediction but also in efficiently handling the high-dimensional data often encountered in clinical datasets. High-dimensional data may contain irrelevant or redundant features that can obscure meaningful patterns, lead to overfitting, and unnecessarily increase computational complexity. Therefore, it is essential to apply feature extraction and selection techniques to reduce the dimensionality of the dataset while retaining the most relevant information for accurate diagnosis [1].

Feature selection is a crucial step in the development of ML models for heart disease diagnosis. It involves identifying the most informative features that contribute to the prediction of heart disease, thereby improving model interpretability, reducing overfitting, and enhancing computational efficiency [2]. Techniques such as Recursive Feature Elimination (RFE), Principal Component Analysis (PCA), and various filter-based methods have been widely applied in medical datasets to improve diagnostic accuracy [3, 4].

The use of ML models such as K-Nearest Neighbor (KNN), Decision Tree Classifier (DTC), Support Vector Machines (SVM), Logistic Regression (LR), and Random Forest (RF) has shown promise in heart disease prediction [5]. However, the performance of these models is highly dependent on the selection of appropriate features. Studies have demonstrated that feature selection not only enhances model performance but also aids in identifying key biomarkers associated with heart disease [6, 7].

Recent research has emphasized the importance of evaluating ML models using robust metrics like the Confusion Matrix and the Receiver Operating Characteristic (ROC) curve. These metrics provide insights into the predictive capabilities of the models and help in comparing different feature selection techniques [8, 9]. The application of these techniques to datasets such as those from the UCI Machine Learning Repository, which includes both continuous and categorical features, has yielded significant improvements in diagnostic accuracy [10].

In this study, we explore various feature selection methods and their impact on the performance of several ML models in diagnosing heart disease. By applying these techniques to a dataset from the UCI repository, we aim to identify the most critical features that contribute to accurate predictions and to evaluate the models' performance using confusion matrices and ROC curves. This research highlights the importance of feature selection in optimizing the diagnostic capabilities of ML models and provides insights for future applications in clinical practice.

II. DATASET

The Heart Disease dataset from the UCI Machine Learning Repository is a well-known resource for studying the classification of heart disease presence in patients based

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24SEP1684

on a set of medical attributes. The dataset comprises 303 instances, each with 14 attributes, including both input features and the target variable. These attributes provide a range of patient information, such as age, gender, chest pain type, resting blood pressure, serum cholesterol level, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise-induced angina, ST depression induced by exercise, the slope of the peak exercise ST segment, number of major vessels colored by fluoroscopy, and thalassemia status. The target variable indicates whether the patient has heart disease (1) or not (0).

The dataset is designed for the classification problem, where the aim is to predict the presence of heart disease using the provided features. It includes both categorical and continuous variables, making it suitable for a variety of data preprocessing and modeling techniques. Researchers and practitioners commonly use this dataset to develop and evaluate machine learning algorithms such as Logistic Regression, Decision Trees, Support Vector Machines, K-Nearest Neighbors, and ensemble methods like Random Forests.

With no missing values and a mix of feature types, this dataset is an excellent tool for testing the effectiveness of various feature selection and machine learning techniques. It serves as a standard benchmark in the field of medical data analysis, particularly in predicting heart disease. The dataset [11] is publicly accessible through the UCI Machine Learning Repository.

The dataset contains one duplicate value as shown in figure 1 below, which we decided to drop.

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
164	38	1	2	138	175	0	1	173	0	0.0	2	4	2	1

Fig 1: Duplicate Record Identified

The statistical summary provides an overview of the dataset's characteristics, highlighting the variability and distribution of each attribute. It reveals a predominantly male sample, with a significant proportion of patients diagnosed with heart disease. There is substantial variability in medical attributes such as cholesterol, blood pressure, and heart rate, all of which are crucial for predicting heart disease. Following is the observations on the Dataset's Statistical Summary.

➢ *Sample Size:*

- The dataset contains 302 instances for each attribute, indicating one missing value from the initial count of 303 in the dataset.

➢ *Age:*

- The age of patients ranges from 29 to 77 years, with a mean age of approximately 54.4 years. The standard deviation is 9.05, indicating a moderate spread of ages around the mean.
- The distribution suggests that most patients are middle-aged or older, with 50% of the data falling between 48 and 61 years.

➢ *Sex:*

- The sex attribute is binary, where 0 represents female and 1 represents male. The mean value of 0.68 suggests that a higher proportion of the patients are male.
- The minimum value is 0, and the maximum value is 1, confirming the binary nature of this attribute.

➢ *Chest Pain Type (cp):*

- The chest pain type (cp) is a categorical variable with values ranging from 0 to 3, indicating different types of chest pain.
- The mean value is approximately 0.96, with a standard deviation of 1.03, suggesting that most patients experience less severe chest pain types.

➢ *Resting Blood Pressure (trtbps):*

- The resting blood pressure values range from 94 to 200 mm Hg, with a mean of about 131.6 mm Hg. This indicates that the majority of patients have relatively high blood pressure, with the standard deviation at 17.56 mm Hg.

➢ *Serum Cholesterol (chol):*

- Cholesterol levels in the dataset vary widely, from 126 to 564 mg/dl, with a mean of 246.5 mg/dl. The standard deviation is 51.75 mg/dl, indicating substantial variability in cholesterol levels among patients.
- The high maximum value suggests that some patients have significantly elevated cholesterol levels.

➢ *Fasting Blood Sugar (fbs):*

- Fasting blood sugar is another binary attribute, where 0 indicates a level ≤120 mg/dl and 1 indicates a level >120 mg/dl.
- The mean of 0.149 suggests that only a small fraction of patients have fasting blood sugar levels above the

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24SEP1684

threshold, indicating most patients do not have elevated fasting blood sugar.

➢ *Resting Electrocardiographic Results (restecg):*

- Restecg has a mean of 0.53, with a standard deviation almost equal to its mean, reflecting a varied distribution across the three categories (0, 1, and 2).

➢ *Maximum Heart Rate Achieved (thalachh):*

- The maximum heart rate achieved by patients ranges from 71 to 202 bpm, with an average of approximately 149.6 bpm. The standard deviation of 22.90 bpm suggests moderate variability in the maximum heart rates.

➢ *Exercise-Induced Angina (exng):*

- Exercise-induced angina is a binary variable, where 0 indicates no angina and 1 indicates the presence of angina.
- The mean value of 0.328 indicates that approximately 32.8% of the patients experienced angina induced by exercise.

➢ *ST Depression Induced by Exercise (oldpeak):*

- The oldpeak attribute, which represents ST depression, ranges from 0 to 6.2 mm with a mean of 1.04 mm. This indicates that most patients experience some level of ST depression, with a few outliers showing significant depression.

➢ *Slope of the Peak Exercise ST Segment (slp):*

- Slope values range from 0 to 2, with a mean of 1.40, suggesting a higher occurrence of upsloping or flat ST segments among patients.

➢ *Number of Major Vessels Colored by Fluoroscopy (caa):*

- The caa attribute, representing the number of major vessels (0 to 3) colored by fluoroscopy, has a mean of 0.719 and a maximum value of 4, suggesting that most patients have few or no major vessels colored.

➢ *Thalassemia (thall):*

- Thall is a categorical variable with values ranging from 0 to 3, and a mean value of approximately 2.31. This indicates a higher occurrence of abnormal results, with most patients likely having some form of thalassemia-related defect.

➢ *Output:*

- The target variable output is binary, indicating the presence (1) or absence (0) of heart disease. The mean of 0.54 suggests that around 54% of the patients in the dataset have been diagnosed with heart disease.

III. EXPLORATORY DATA ANALYSIS

The Heart Disease dataset provides valuable insights into the distribution and characteristics of heart disease among patients. This analysis seeks to understand the relationship between gender and the presence of heart disease, as well as the distribution of different types of chest pain among males and females.

➢ *Question 1: What is the proportion of males and females having heart disease or not?*

The dataset comprises 302 instances, with 138 instances indicating no heart disease (0) and 164 instances indicating the presence of heart disease (1). It is observed that 68.2% of the patients are male, while 31.8% are female. Among the males, 44.7% are more likely to suffer from heart disease, whereas 55.3% are less likely. For females, the scenario is different, with 75% having a higher chance of heart disease and only 25% having a lower chance.

➢ *Question 2: What is the proportion of males and females experiencing different types of chest pain?*

Chest pain is categorized into four types: typical angina, non-anginal pain, atypical angina, and asymptomatic. Among the male patients, approximately 50.5% experience chest pain typical of angina, 24.8% suffer from non-anginal pain, 15.5% have atypical angina, and 9.2% are asymptomatic. For female patients, the distribution is slightly different, with 40.6% experiencing typical angina, 36.5% suffering from non-anginal pain, 18.8% experiencing atypical angina, and 4.2% being asymptomatic. Refer figure 2.

The data highlights gender differences in both the likelihood of heart disease and the types of chest pain experienced. Males tend to have a lower likelihood of heart disease than females, who are more likely to suffer from the condition. Additionally, typical angina is the most common type of chest pain in both genders, but its prevalence is higher in males. Non-anginal pain is also common, especially in females, where it accounts for 36.5% of the cases. The patterns observed in this exploratory data analysis provide a foundation for further investigation into the factors influencing heart disease and the effectiveness of different diagnostic approaches.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24SEP1684

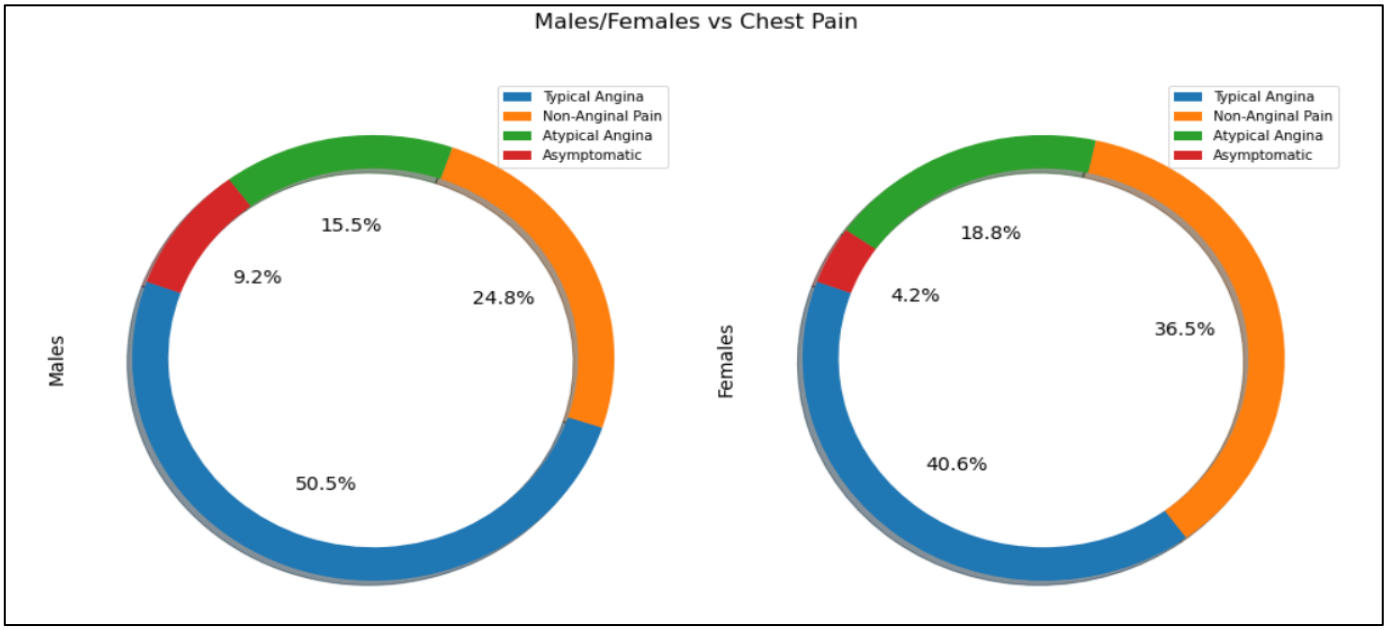

Fig 2: Males/Females Vs Chest Pain

➢ *Question 3: What is the frequency and proportion of patients with heart disease based on the type of chest pain?*

The analysis also explores the relationship between the type of chest pain experienced by patients and the presence or absence of heart disease. The types of chest pain include typical angina, non-anginal pain, atypical angina, and asymptomatic cases. The frequency and proportion of patients diagnosed with heart disease or not, categorized by these chest pain types, provide insights into the predictive value of chest pain in diagnosing heart disease.

Figure 3 illustrates these relationships, showing how different chest pain types are associated with varying likelihoods of heart disease. For instance, patients with typical angina may show a higher proportion of heart disease cases, whereas those with non-anginal pain might have a different distribution. Understanding these patterns can help in better assessing the risk of heart disease based on presenting symptoms.

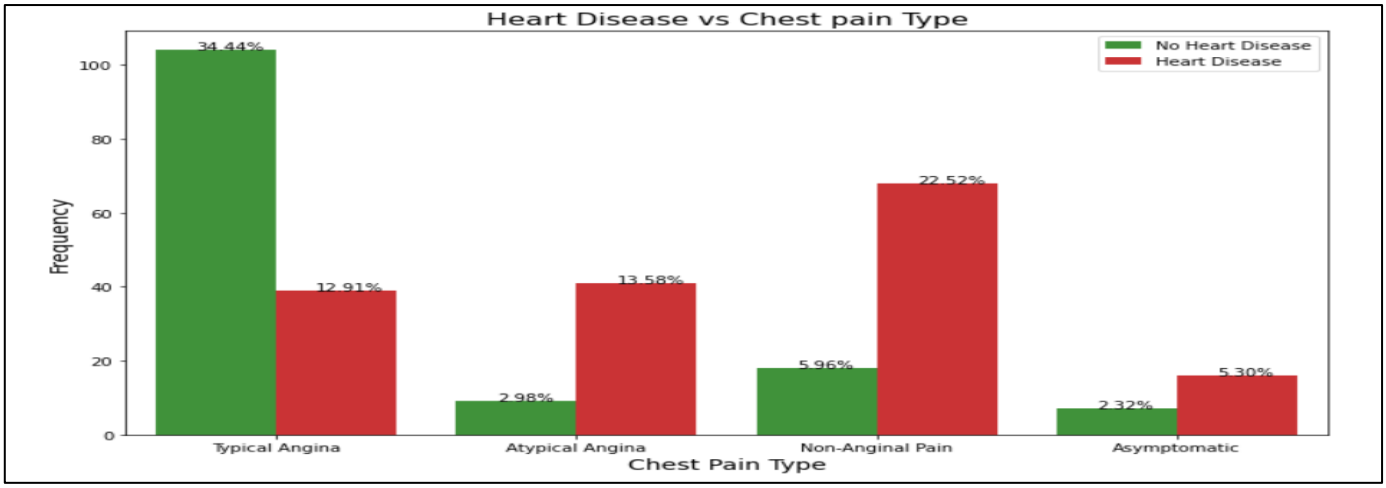

Fig 3: Heart Disease Vs Chest Pain Type

➢ *Question 4: What is the association between cholesterol levels and age based on the sex of the patients?*

This analysis examines the relationship between cholesterol levels and age, with a focus on differences between male and female patients. Total cholesterol levels less than 200 milligrams per deciliter (mg/dL) are considered desirable for adults. The findings indicate that both male and female patients tend to have similar cholesterol levels up to the age of 60. However, among patients older than 60, some female patients exhibit higher cholesterol levels compared to their male counterparts of the same age, as illustrated in Figure 4. This observation suggests a potential gender-based difference in cholesterol levels as patients age, which could have implications for assessing heart disease risk in older adults.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24SEP1684

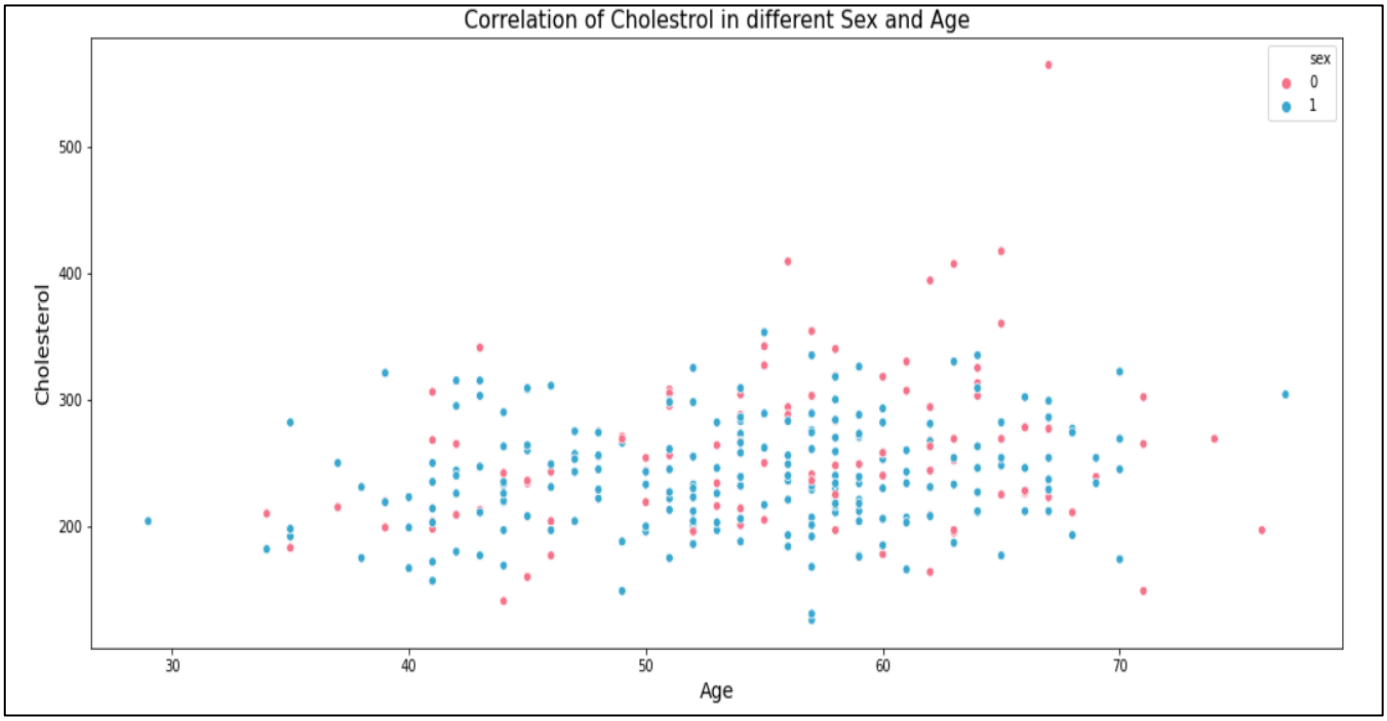

Fig 4: Correlation of Cholestrol wrt Different Sex and Age

IV. FEATURE SELECTION

Feature selection is a crucial step in building predictive models, aimed at enhancing model performance and interpretability by reducing the number of input variables. This process involves selecting a subset of relevant features from a larger set, thereby minimizing redundancy and mitigating the risk of overfitting [1]. Effective feature selection not only improves the model's accuracy but also its computational efficiency.

In this paper, we explore two prominent feature selection techniques: wrapper methods and Random Forest-based selection. Wrapper methods assess the usefulness of feature subsets by training and testing a model on various feature combinations and selecting the best-performing subset based on model accuracy [12]. While these methods can capture feature interactions and optimize performance, they can be computationally intensive due to multiple model evaluations.

In contrast, Random Forest-based feature selection uses the importance scores derived from an ensemble of decision trees to identify significant features [13]. This method evaluates the contribution of each feature to the reduction in prediction error across the trees in the forest, as illustrated in Figure 5. It is advantageous for its ability to handle large datasets and complex feature interactions while maintaining computational efficiency.

Using Random Forest-based feature selection, we identified 8 significant features with a threshold of 0.08. The selected features are: 'age', 'cp', 'trtbps', 'chol', 'thalachh', 'oldpeak', 'caa', and 'thall'. In contrast, two exhaustive feature selection methods, EFS1 and EFS2, were applied to evaluate feature subsets. EFS1 achieved a performance score of 85.12%, selecting the features: 'sex', 'cp', 'trtbps', 'restecg', 'exng', 'oldpeak', 'slp', 'caa', and 'thall'. EFS2, with a performance score of 83.88%, identified the following features: 'sex', 'cp', 'trtbps', 'exng', 'slp', 'caa', and 'thall'.

Given the results, we opted for EFS1 as our feature selection method due to its superior performance score and comprehensive feature subset selection. This choice ensures that our model benefits from a robust and accurate set of features, thereby enhancing its predictive capability.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24SEP1684

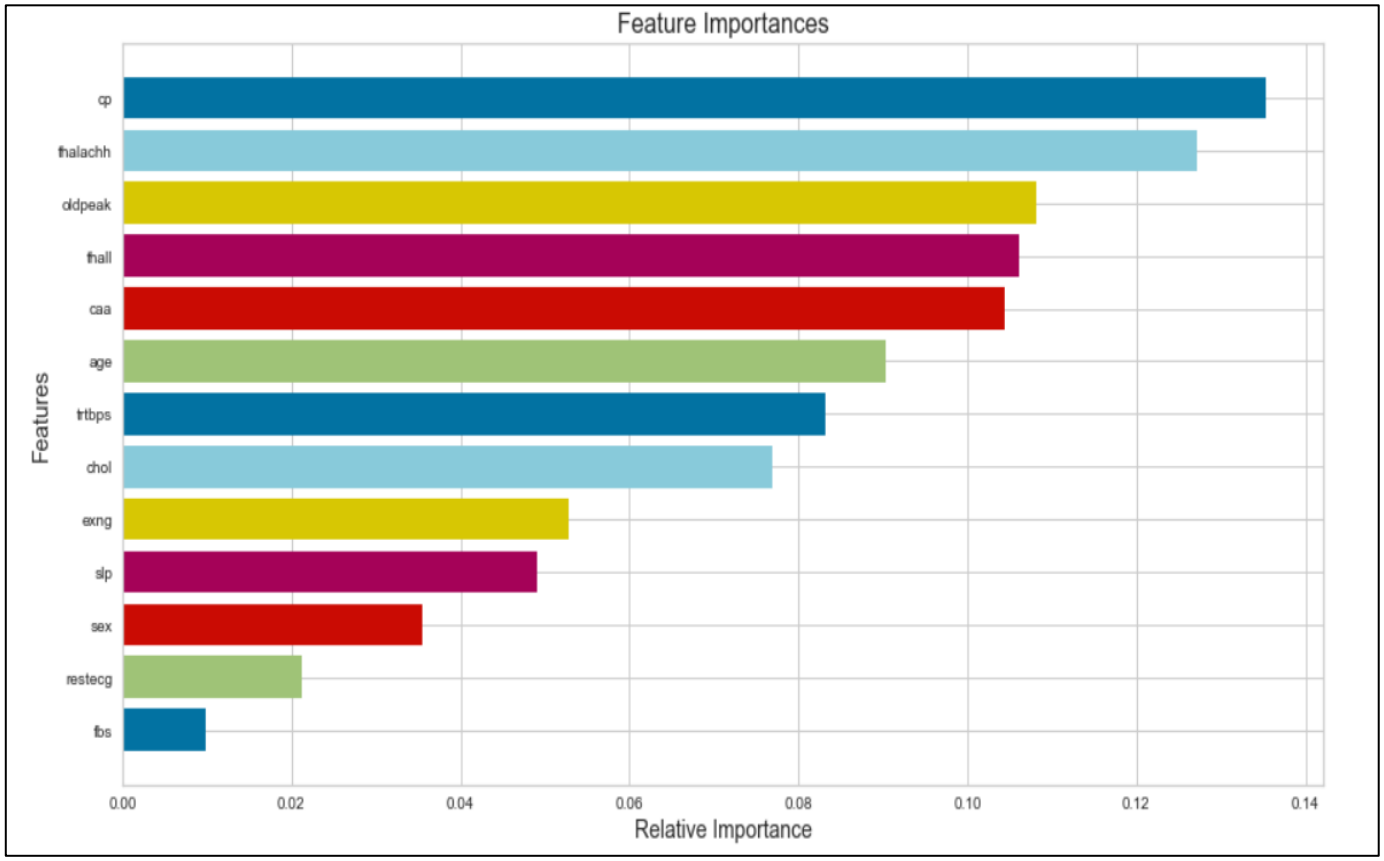

Fig 5: Feature Importance

V. MACHINE LEARNING MODELS

In this study, we applied several machine learning algorithms to evaluate their performance on the heart disease dataset, after splitting the data into an 80:20 ratio for training and testing and applying standardization using Standard Scaler. The models evaluated include Logistic Regression, Decision Tree, Random Forest, K Nearest Neighbour (KNN), and Support Vector Machine (SVM).

- **Logistic Regression** is a fundamental classification algorithm that models the probability of a binary outcome based on one or more predictor variables [14]. Our implementation achieved an accuracy of 84.62%, demonstrating its effectiveness in distinguishing between the presence and absence of heart disease.
- **Decision Tree** is a versatile model that uses a tree-like structure of decisions and their possible consequences [15]. It offers interpretability and handles both categorical and numerical data. In our case, the Decision Tree model had an accuracy of 76.92%, which indicates a lower performance compared to other models, potentially due to overfitting or insufficient tree depth.
- **Random Forest** is an ensemble method that combines multiple decision trees to improve predictive performance and control overfitting [13]. This model achieved an accuracy of 81.32%, reflecting its robust performance in handling complex interactions between features but still falling short of the top-performing models in this evaluation.
- **K Nearest Neighbour (KNN)** is a non-parametric method that classifies data points based on the majority label of their nearest neighbors [16]. As illustrated in Figure 6, the optimal value for K was determined to be 12, which provided the best error rate. The KNN model with K=12 achieved the highest accuracy of 89.01%, making it the most effective model in our analysis.
- **Support Vector Machine (SVM)** is a powerful classification technique that finds the optimal hyperplane to separate different classes [17]. Our SVM model achieved an accuracy of 85.71%, showcasing strong performance and ability to generalize well to unseen data.

Based on the accuracy metrics, the KNN model with K=12 demonstrated the highest performance among the evaluated algorithms. Therefore, KNN is suggested as the best model for this dataset due to its superior accuracy and effective handling of the classification task.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24SEP1684

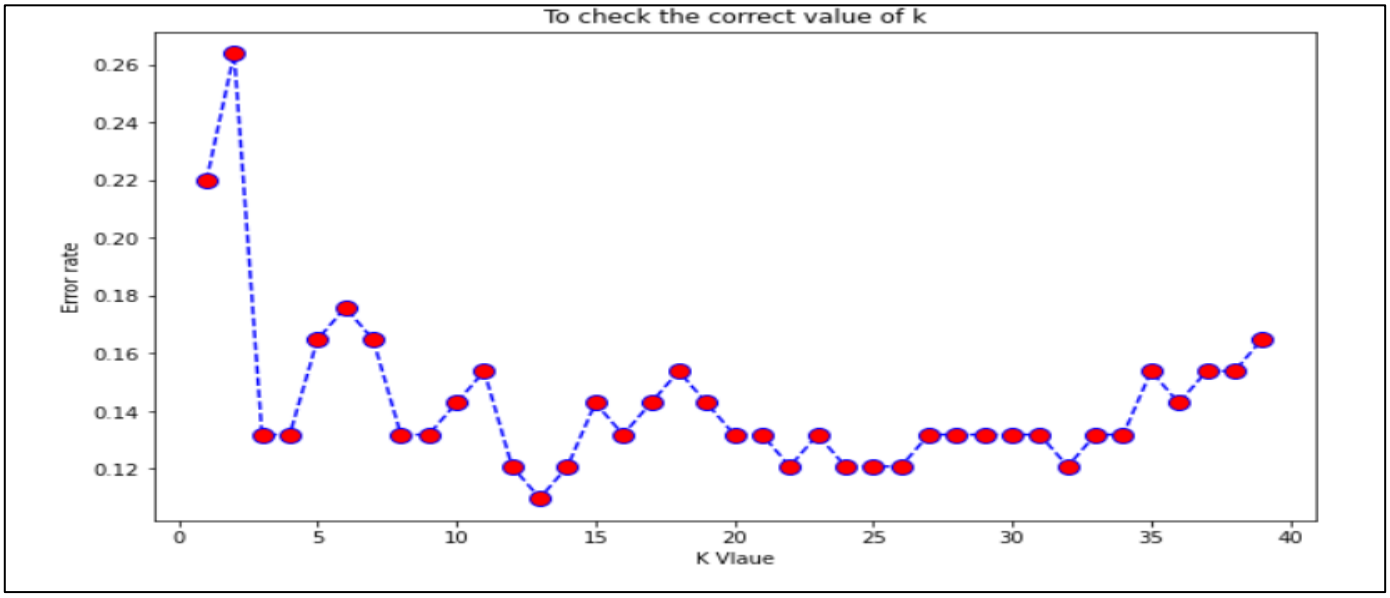

Fig 6: To Obtain Appropriate Value of K

VI. CONCLUSION

This study aimed to enhance heart disease diagnosis through the optimization of feature selection techniques and evaluation of various machine learning models. By leveraging the Heart Disease dataset from the UCI Machine Learning Repository, we applied feature extraction and reduction methods to identify the most critical attributes for accurate heart disease prediction. The dataset was split into training and testing sets, and standardization was performed to ensure consistency in the model evaluations.

Feature selection techniques, including wrapper methods and Random Forest-based approaches, played a crucial role in improving model performance. The exhaustive feature selection method EFS1, which identified a subset of features including 'sex', 'cp', 'trtbps', 'restecg', 'exng', 'oldpeak', 'slp', 'caa', and 'thall', achieved the highest performance score of 85.12%, surpassing EFS2's score of 83.88%. This indicates that EFS1 offered a more effective feature subset for enhancing diagnostic accuracy.

When evaluating machine learning models, K-Nearest Neighbour (KNN) with K=12 emerged as the most accurate model, achieving an impressive accuracy of 89.01%. This was followed by Support Vector Machine (SVM) with an accuracy of 85.71%, Logistic Regression at 84.62%, Random Forest at 81.32%, and Decision Tree at 76.92%. The high accuracy of KNN underscores its efficacy in handling the classification task for heart disease diagnosis, making it the most effective model among those evaluated.

The results underscore the significance of feature selection in optimizing model performance. By focusing on the most relevant features, we were able to enhance the diagnostic capabilities of our models, thereby improving their accuracy and reliability. The findings suggest that future research and practical applications in heart disease diagnosis should prioritize effective feature selection techniques to maximize the performance of machine learning models.

Overall, this study provides valuable insights into the application of machine learning for heart disease diagnosis and highlights the importance of feature selection in developing accurate and efficient predictive models. Future work could explore additional feature selection techniques and machine learning algorithms to further refine diagnostic tools and contribute to better healthcare outcomes.

Feature selection is a crucial step in building predictive models, aimed at enhancing model performance and interpretability

REFERENCES

[1]. Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*, 1157-1182.
[2]. Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering, 40*(1), 16-28.
[3]. Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence, 97*(1-2), 273-324.
[4]. Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374*(2065), 20150202.
[5]. Shouman, M., Turner, T., & Stocker, R. (2012). Applying k-nearest neighbour in diagnosing heart disease patients. *International Journal of Information and Education Technology, 2*(3), 220-223.
[6]. Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J. J., Sandhu, S., ... & Guppy, K. H. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American Journal of Cardiology, 64*(5), 304-310.
[7]. Duda, R. O., Hart, P. E., & Stork, D. G. (2001). Pattern classification. *John Wiley & Sons*.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24SEP1684

[8]. Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*(1), 29-36.
[9]. Powers, D. M. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, *2*(1), 37-63.
[10]. Dua, D., & Graff, C. (2017). UCI machine learning repository. *University of California, Irvine, School of Information and Computer Sciences*.
[11]. Dua, D., & Graff, C. (2019). UCI Machine Learning Repository [Heart Disease Data Set]. Irvine, CA: University of California, School of Information and Computer Science. Available from: https://archive.ics.uci.edu/ml/datasets/Heart+Disease.
[12]. Kohavi, R., & John, G. H. (1997). Wrappers for Feature Subset Selection. *Artificial Intelligence*, 97(1-2), 273-324.
[13]. Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32.
[14]. Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression*. Wiley. doi:10.1002/9781118548387
[15]. Quinlan, J. R. (1986). Induction of Decision Trees. *Machine Learning*, 1(1), 81-106. doi:10.1007/BF00116251
[16]. Cover, T. M., & Hart, P. E. (1967). Nearest Neighbor Pattern Classification. *IEEE Transactions on Information Theory*, 13(1), 21-27. doi:10.1109/TIT.1967.1053964
[17]. Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, 20(3), 273-297. doi:10.1007/BF00994018