Heart Disease Prediction using Machine Learning

Reshma Kanna R [#1]
School of Advanced Sciences
Vellore Institute of Technology
Vellore, India
reshmakanna22@gmail.com

Venkataramana B [#2]
School of Advanced Sciences
Vellore Institute of Technology
Vellore, India
venkataramana.b@vit.ac.in

***Abstract*— This study focuses on the development of a heart disease prediction model employing machine learning techniques. The dataset comprises critical variables, including age, gender, chest pain type, blood pressure, cholesterol levels, and various cardiovascular indicators. The procedural steps involve thorough exploration of the dataset, data visualization to discern patterns, and feature engineering encompassing feature selection. Data preprocessing strategies address categorical features and ensure proper feature scaling. The model building phase incorporates three classifiers, namely K Neighbours, Decision Tree, and Random Forest, each evaluated for their predictive performance. This project contributes to predictive healthcare analytics by leveraging a diverse set of features to enhance the accuracy and interpretability of heart disease predictions.**

Keywords – Heart disease prediction, Machine learning techniques, K Neighbours Classifier, Decision Tree Classifier, Random Forest Classifier, Dataset exploration, Cardiovascular indicators.

I. INTRODUCTION

Heart disease remains the leading cause of mortality globally, with significant impacts on public health and healthcare systems. According to the World Health Organization (WHO), cardiovascular diseases, including heart attacks and strokes, account for a substantial portion of global deaths, representing 31% of all mortalities. The alarming prevalence of heart-related ailments underscores the urgency for enhanced diagnostic tools and predictive models to mitigate the escalating mortality rates associated with these conditions.

This research project focuses on the utilization of machine learning techniques for heart disease prediction, employing a Python-based approach encompassing data visualization, feature engineering, data preprocessing, and model building. The dataset utilized in this study comprises various demographic, clinical, and physiological attributes, providing a comprehensive representation of factors influencing heart disease incidence.

Understanding the common types of heart diseases is essential for effective prevention, diagnosis, and management strategies. Shown below in this table is a brief overview of some prevalent types of heart diseases observed today:

TABLE I. An Overview of Common Heart Diseases

Heart Disease	Description
Arrhythmia	Improper heartbeat characterized by irregularity, slowness, or rapidity.
Cardiac Arrest	Sudden loss of heart function, leading to loss of consciousness and cessation of breathing.
Congestive Heart Failure	Chronic condition where the heart fails to pump blood effectively.
Congenital Heart Disease	Abnormalities in the heart's structure present at birth.
Coronary Artery Disease	Disease affecting the major blood vessels supplying the heart, leading to reduced blood flow.
High Blood Pressure	Condition where blood exerts excessive force against artery walls.
Peripheral Artery Disease	Narrowing of blood vessels, reducing blood flow to the limbs.
Stroke	Interruption of blood supply to the brain, leading to brain damage.

By systematically evaluating and comparing the performance of different machine learning algorithms, this research aims to identify the most effective model for heart disease prediction. The insights gleaned from this study have the potential to inform clinical decision-making processes, enabling healthcare practitioners to implement timely interventions and preventive measures for individuals at risk of heart disease.

II. LITERATURE REVIEW

With the advancement of technology, particularly in the realm of machine learning, there is growing interest in leveraging computational techniques to enhance the prediction and diagnosis of cardiovascular diseases (CVDs). This literature review aims to provide a comprehensive overview of recent research endeavours focused on utilizing machine learning algorithms for heart disease prediction.

The paper “Heart Disease Prediction using Machine Learning Techniques” by Pooja Anbuselvan investigates the effectiveness of various machine learning algorithms in predicting heart disease. Through analyzing supervised learning models like Logistic Regression, Naïve Bayes, Support Vector Machine, K-Nearest Neighbors, Decision Tree, Random Forest, and XGBoost, the study aims to identify the most accurate predictor. Results reveal Random Forest as the top performer with 86.89% accuracy, while K-Nearest Neighbor fares the poorest at 57.83%. In "Heart Disease Prediction Using Machine Learning Techniques" by Apurv Garg et al., machine learning is explored for detecting cardiovascular diseases (CVDs), a major global health concern. The study employs K-Nearest Neighboursr (KNN) and Random Forest algorithms to classify individuals based on attributes like chest pain, cholesterol level, and age, achieving prediction accuracies of 86.885% and 81.967% respectively. While acknowledging limitations, the research underscores the significance of machine learning in predicting heart disease and its potential to improve healthcare outcomes. The authors anticipate further advancements in machine learning to enhance healthcare applications. Authors V.V. Ramalingam, Ayantan Dandapath, and M Karthik Raja present a comprehensive survey on machine learning techniques for heart disease prediction. The study evaluates supervised learning algorithms including Naïve Bayes,

Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree, Random Forest, and ensemble models. While SVM consistently achieves high accuracy, the performance of Decision Trees varies due to potential overfitting. The survey highlights the potential of machine learning in heart disease prediction but emphasizes the need for further research to address challenges such as handling high-dimensional data and optimizing algorithm ensembles for improved accuracy.

III. DATA

The Heart Disease dataset has been procured from UC Irvine Machine Learning Repository, a popular inventory of databases for the purpose of usage in research papers and projects. The dataset contains 14 variables and 303 instances related to various physiological parameters and medical indicators, aimed at predicting the presence or absence of heart disease in individuals.

A. Data Exploration

In this exploration of the Heart Disease dataset, we began by examining its structure and characteristics. Initially, we determined the dataset's size and shape, then we identified the column headers to understand the available features within the dataset. By inspecting the tail end of the dataset, we gained a glimpse into its contents, which can help identify any potential patterns or anomalies. Detecting missing values is essential, as it allows for appropriate handling to ensure data integrity. Furthermore, obtaining basic information about the dataset, such as non-null counts and data types, provides a comprehensive overview necessary for subsequent analysis. Descriptive statistics offer numerical summaries of the dataset's features, including measures like mean, standard deviation, and quartile values, aiding in understanding the distribution and characteristics of the data for predictive modelling endeavours.

B. Data Preprocessing

In preparation for machine learning model training, several preprocessing steps were applied to the Heart Disease dataset. Categorical variables, including 'sex', 'cp' (chest pain type), 'fbs' (fasting blood sugar), 'restecg' (resting electrocardiographic results), 'exang' (exercise-induced angina), 'slope' (slope of the peak exercise ST segment), 'ca' (number of major vessels colored by fluoroscopy), and 'thal' (thalassemia), underwent conversion into dummy variables using the get_dummies() function from the Pandas library. Subsequently, the dataset was split into dependent features (X) and the target variable (y), where 'target' typically indicates the presence or absence of heart disease.

These preprocessing steps collectively prepared the dataset for machine learning model training by transforming categorical variables, scaling numerical features, and separating independent variables from the target variable.

C. Feature Engineering

Feature engineering involves selecting relevant features for modeling. A correlation matrix is computed for all features, and the top correlated features are extracted. Furthermore, numeric features were standardized using the StandardScaler from scikit-learn. Standardization ensures that all features are on the same scale, which is essential for various machine learning algorithms, particularly those relying on distance metrics or gradient descent.

This process guides the selection of features to include or exclude in the predictive model, aiding in the development of more accurate heart disease prediction models.

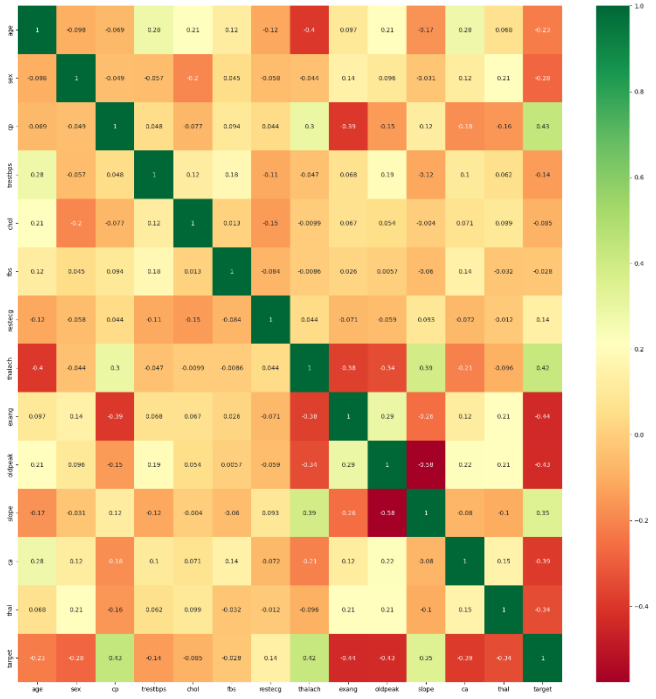

Fig. 1. Correlation heatmap of all the study parameters

D. Data Visualization

In this section, data visualization techniques were employed to gain insights into the Heart Disease dataset. The primary libraries utilized include Matplotlib and Seaborn for plotting purposes.

Firstly, a histogram was constructed to provide a comprehensive overview of the dataset's distribution. The entire dataset was visualized using histograms, with each feature represented to showcase their respective distributions. This visualization aids in understanding the spread and variability of each feature within the dataset.

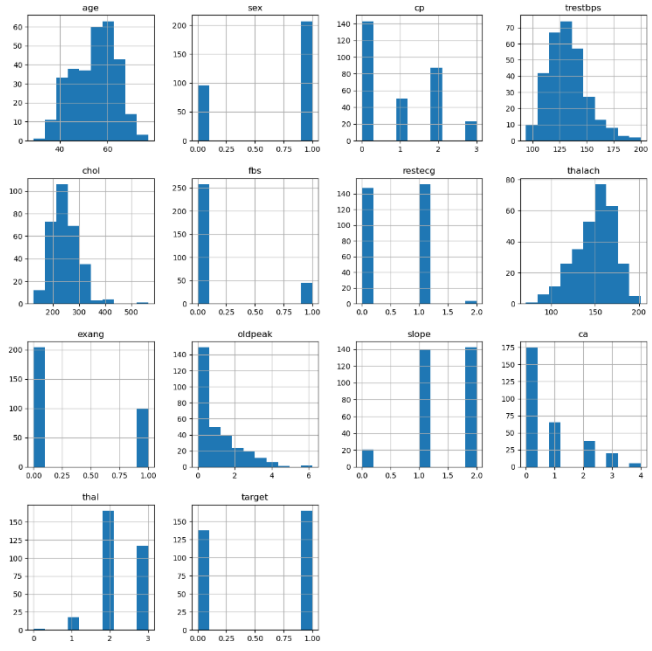

Fig. 2.1 Histogram of all the study parameters

Subsequently, a count plot was generated, illustrating the distribution of target values within the dataset. This visualization offers insights into the class distribution and helps ascertain whether the dataset is balanced or skewed towards certain outcomes. The x-axis denotes the target variable, while the y-axis represents the count of occurrences.

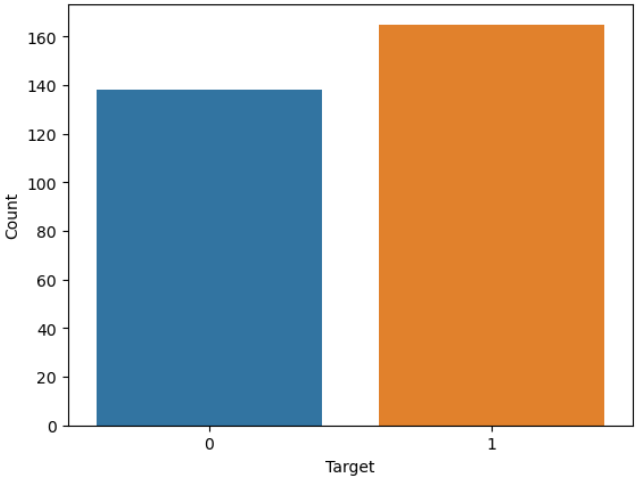

Fig. 2.2. Count plot of target variable

IV. METHODOLOGY

As stated earlier, we employed a machine learning approach to predict heart disease utilizing three distinct algorithms: k-Nearest Neighbours (KNN), Random Forest (RF), and Decision Trees (DT). We then trained and evaluated each machine learning model using rigorous cross-validation techniques to assess their performance in terms of accuracy. Finally, we compared the results obtained from the three models to determine their effectiveness in predicting heart disease.

A. Model Building

1) KNN:

The k-Nearest Neighbors (KNN) algorithm, a non-parametric, instance-based learning method, was employed for heart disease prediction. KNN operates by identifying the 'k' nearest data points in the feature space to a given point of interest and assigning the majority class label among these neighbors. We utilized scikit-learn's KNeighborsClassifier for model development, optimizing its performance through cross-validation by iterating over 'k' values to find the one yielding the highest cross-validation score, which was found to be 12. Following determination of the optimal 'k' value, we trained the KNN classifier on preprocessed data. Through this process, the model learned patterns within the data, establishing relationships between feature variables and the target variable - presence or absence of heart disease.

2) Decsion Tree:

The Decision Tree classifier is a versatile supervised learning algorithm utilized for classification and regression tasks. In our research, we employed scikit-learn's DecisionTreeClassifier, iterating over various maximum depth values to optimize model performance through cross-validation. This process enables the algorithm to recursively partition the feature space, establishing decision rules based on feature values to classify individuals as at risk or not at risk of heart disease. The optimal maximum depth value, determined through cross-validation was found to be 3 and it controls the tree's complexity, preventing overfitting.

3) Random Forest:

The Random Forest classifier, a powerful ensemble learning method, is utilized for classification tasks such as heart disease prediction. This algorithm constructs multiple decision trees during training and outputs the mode or mean prediction of the individual trees, reducing overfitting and variance. In our research, scikit-learn's RandomForestClassifier was employed, with the number of estimators optimized through cross-validation. After determining the optimal number of estimators, which was found to be 90, the classifier was trained on preprocessed data containing relevant features. The model aggregates predictions from individual trees, leveraging their collective wisdom to make accurate predictions.

B. Model Evaluation

In assessing the efficacy of machine learning models for heart disease prediction, our study employed accuracy as the primary evaluation metric. Accuracy represents the proportion of correctly classified instances among all instances in the dataset, providing a comprehensive measure of a model's overall predictive performance. A higher accuracy score indicates that the model effectively discerns between individuals at risk of heart disease and those who are not, thereby facilitating more accurate risk assessment and diagnosis. By evaluating accuracy across multiple classifiers such as k-Nearest Neighbours(KNN), Decision Tree, and Random Forest, our study provides valuable insights into the relative strengths and weaknesses of each approach.

V. RESULTS

In evaluating the effectiveness of machine learning models for heart disease prediction, our study focused on three classifiers: k-Nearest Neighbors (KNN), Decision Tree, and Random Forest. Each classifier was trained on a comprehensive dataset containing details like medical history and lifestyle factors associated with heart disease. The evaluation results highlighted distinct performance metrics across the classifiers. The KNN classifier demonstrated an accuracy rate of 83.94% with a k value set to 12, showcasing its ability to accurately classify individuals based on their susceptibility to heart disease. Meanwhile, the Decision Tree classifier, restricted to a maximum depth of 3, achieved a slightly lower accuracy score of 82.46%, yet still provided reliable predictions.

Of significant note, the Random Forest classifier emerged as the top performer, boasting the highest accuracy rate of 86.89%. Comprised of an ensemble of 90 decision trees, the Random Forest model exhibited superior predictive capabilities by leveraging the collective insights from its constituent trees to discern intricate patterns within the heart disease dataset. These findings underscore the potential of ensemble methods, particularly Random Forest, in enhancing predictive accuracy for heart disease prediction tasks.

Overall, our study highlights the importance of evaluating multiple machine learning algorithms to identify the most effective approach for heart disease prediction, with Random Forest showing promise as a robust model for risk assessment and diagnosis in clinical settings.

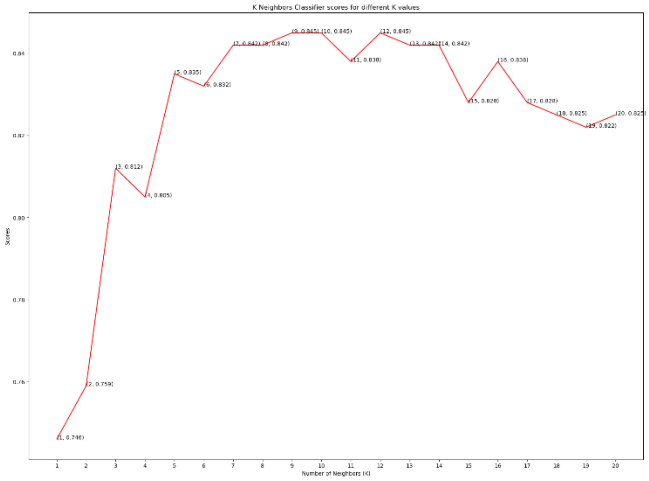

Fig. 3.1. K Neighbours Classifier scores for different K values.

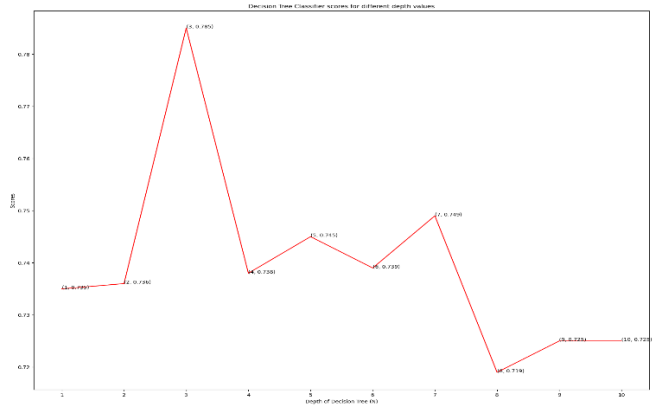

Fig. 3.2. Decision Tree Classifier scores for different depth values.

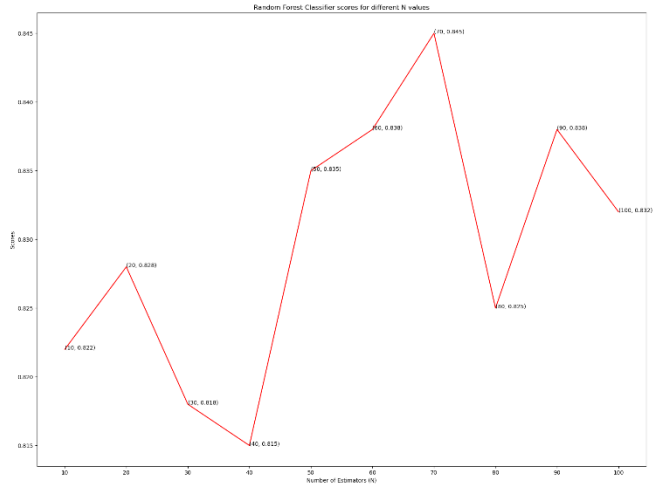

Fig. 3.3. Random Forest Classifier scores for different N values.

VI. REFERENCES

[1] Rubini PE, Dr.C.A.Subasini, Dr.A.Vanitha Katharine, V.Kumaresan, S.GowdhamKumar, T.M. Nithya, "A Cardiovascular Disease Prediction using Machine Learning Algorithms", Annals of R.S.C.B., 2021.

[2] Chintan M. Bhatt, Parth Patel, Tarang Ghetia and Pier Luigi Mazzeo, "Effective Heart Disease Prediction Using Machine Learning Techniques" , repository: https://doi.org/10.3390/a16020088, 2023.

[3] M. Marimuthu, M. Abinaya, K. S. Hariesh, K. Madhankumar, V. Pavithra, "A Review on Heart Disease Prediction using Machine Learning and Data Analytics Approach", International Journal of Computer Applications, 2018.

[4] Apurv Garg, Bhartendu Sharma and Rijwan Khan, "Heart disease prediction using machine learning techniques", IOP Conference Series: Materials Science and Engineering, 2020.

[5] Pooja Anbuselvan, "Heart Disease Prediction using Machine Learning Techniques", International Journal of Engineering Research & Technology, 2020.

[6] V.V. Ramalingam, Ayantan Dandapath, M Karthik Raja, "Heart disease prediction using machine learning techniques: a survey", International Journal of Engineering & Technology, 2018.

[7] Jaymin Patel, Prof.TejalUpadhyay, Dr. Samir Patel, "Heart Disease Prediction Using Machine learning and Data Mining Technique", repository: www.csjournalss.com.

[8] Abhijeet Jagtap, Priya Malewadkar, Omkar Baswat, Harshali Rambade, "Heart Disease Prediction using Machine Learning", International Journal of Research in Engineering, Science and Management, 2019.

[9] Harshit Jindal, Sarthak Agrawal, Rishabh Khera, Rachna Jain and Preeti, Nagrath, "Heart disease prediction using machine learning algorithms", IOP Conf. Series: Materials Science and Engineering, 2020.

[10] Baban.U. Rindhe, Nikita Ahire, Rupali Patil, Shweta Gagare, Manisha Darade, "Heart Disease Prediction Using Machine Learning", International Journal of Advanced Research in Science, Communication and Technology (IJARSCT), 2021.