Heart Disease Prediction using Wrapper's Method by Machine learning models

Likhith Edupuganti[1],Sai Chandhan.G[2],K Hasith[3], P Balaji[4],N.Harshith[5] and B.Sriram[6]

ABSTRACT

Heart is the next major organ comparing to brain which has more priority in Human Body. It pumps the blood and supplies to all organs of the whole body. Day by Day the cases of heart disease are increasing at a rapid rate & its very important and concerning to predict any such diseases beforehand. According to World Health Organization (WHO), cardiovascular diseases (CVD) are the major health concern worldwide and a leading cause of death. CVDs were responsible for 32% of all global deaths in 2019, as estimated by World Health Organization. Heart attacks and strokes were responsible for 85% of these deaths. Some of the machine learning are used to predict the heart disease, such as Logistic Regression, Support Vector Machine (SVM), Decision Tree, Random Forest, K- Nearest Neighbor(KNN) for Prediction. The given heart disease prediction system enhances medical care and reduces the cost. This project provides an insight of the existing algorithm, and it gives us significant knowledge that can help us predict the patients with heart disease.

Keywords

Heart Disease, Machine Learning, World Health Organization (WHO), Cardiovascular Disease (CVD), Heart attack & Strokes.

1. INTRODUCTION

Cardiovascular diseases (CVDs) are the leading cause of death in the world. Each year, around 17.9 million people die from cardiovascular diseases, accounting for 32% of all deaths worldwide, according to the World Health Organization (WHO). CVDs are illnesses that affect the heart and blood vessels, including heart attacks and strokes. Smoking, unhealthy diet and lack of exercise increase your risk of heart disease. Prevention is possible through healthy lifestyle choices, regular check-ups and by predicting the risk before the attack.

Machine Learning (ML) is a very vast and diverse field, and its scope and implementation are increasing day by day. It is a method for extracting and analyzing implicit and explicit data. Machine Learning plays a vital role in predicting heart disease by processing vast health data to uncover patterns and risk factors. The prediction of this disease before being infected is part of the prevention methods. These Machine learning techniques use different types of classifiers, including Supervised, Unsupervised, and Reinforcement Learning, to make predictions more precisely and measure the accuracy of a dataset.

To study this problem and detect cardiac patients, our approach involves identifying individuals based on specific attributes such as Age, Sex(0 = M and 1 = F), Chest Pain (CP . CP 1: Typical angina . CP 2: Atypical angina . CP 3: Non-anginal pain . CP 0: Asymptomatic), resting blood pressure (Normal pressure: 120/80 or lower . Stage 1: 130 to 139 mmHg/80 to 89 mmHg . Stage 2: 140/90 mmHg or higher . Stage 3: 180/120 or Higher), Cholesterol Levels (Normal: <200 mg/dL (5.17 mmol/L) . Borderline High: 200 to 239 mg/dL (5.17 to 6.18 mmol/L) . High: 240 mg/dL (6.21 mmol/L) or greater.), Fasting Blood Sugar (Glucose Test: a common blood test to diagnose prediabetes, diabetes or gestational diabetes.), Resting electrocardiographic measurement (0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria), Thalach (Person's Maximum heart rate achieved HR^{max} – the highest number of beats per minute of the heart), Exang (Exercise induced angina . 1 = yes ; 0 = no), Old Peak (ST depression induced by exercise relative to rest), Slope, CA (number of major vessels [0 - 3] coloured by fluoroscopy), Thal (A Blood disorder called thalassemia [3 = normal; 6 = fixed defect; 7 = reversable defect]) and Target (0 = Absence of Heart Disease & 1 = Presence of Heart Disease).

This project focuses on the following Algorithms like Logistic regression, SVM, Decision tree, Random Forest and KNN. The algorithm that achieves the highest accuracy among 3 to 4 models will be considered the most effective method for prediction. The Objective of this project is to propose the most efficient algorithm within the existing for prediction of Cardiovascular Disease(CVD) patient based on their medical attributes. Ultimately, by establishing the most accurate predictive algorithm, we aim to enhance patient outcomes and contribute to effective prevention strategies for cardiovascular diseases.

figure 1 depicts the parts of human heart such as Left atrium, Tricuspid valve, Aortic valve, Mitral valve, Superior vena Right atrium, Right ventricle, Left ventricle, Aorta, cava and Interior vena cava, Pulmonary vein, Pulmonary valve, Pulmonary artery.

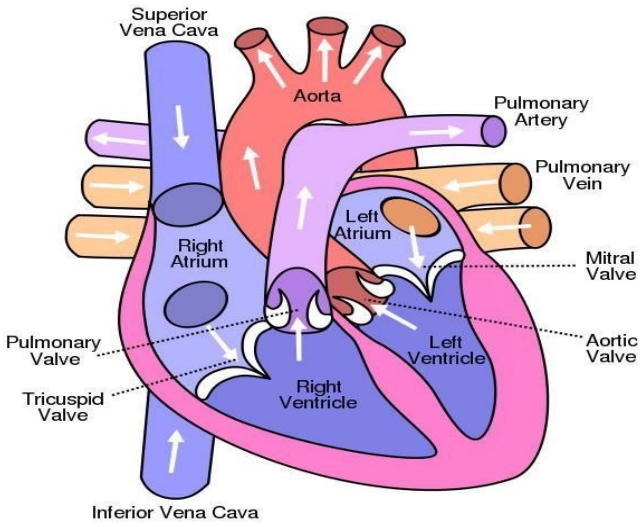


Figure 1 : Human Heart

2. LITERATURE SURVEY

Other researchers works have been evaluated and it is appreciated that the methodology to predict heart disease from the particulars of the patient provided should be the most efficient. In many works, different cutting-edge technologies have been employed which resulted in varying degrees of success; some achieved high-accuracy results while other methods had shortcomings in their predictive capabilities. Here the strengths and

weaknesses of the various methods employed for the prediction of the heart disease are discussed.

Avinash Golande [1] used Machine Learning Techniques to predict Heart Attack effectively. Naive Bayes, decision trees, k-Nearest Neighbours are prevalent machine learning techniques used for classification tasks. However, the precision of each technique may differ based on specific problem and dataset being analysed. In some cases, decision trees may achieve higher accuracy than other algorithms.

Santhana Krishnan [2] used decision trees and Nave Bayes machine learning algorithms, to predict heart attacks. The Decision Tree Model achieved a precision rate of 91% and Nave Bayes' accuracy level was 87%. They presumed that the best algorithm was a decision tree for handling data sets.

Aditi Gavhane et al. [3] worked on machine learning applications by using certain parameters like lifespan, gender, heart rate, etc. to foresee the vulnerability of heart problem. To train and test the dataset, they used neural network supervised algorithms, that is MLP (multilayer perceptron), which gives reliable outcomes from the user's input. Machine learning algorithm using neural network is the most accurate and reliable algorithm.

Rati Goel [4] worked on SVM, LR, RF, Decision tree, Naive Bayes and KNN algorithms of ML and compare their efficiency based on basic parameters like chest pain (CP), gender, blood pressure (BP), and cholesterol to predict heart disease. He concludes that the SVM has better performance in comparison to other machine learning algorithms. Manjula P et al. [5] used various ML techniques like SVM, Decision Tree, Random Forest, Navie Bayes and KNN in their model with Vulnerability factors like age, gender, arterial pressure, cholesterol levels, and family background of heart problem to train and test their models of machine learning. The random forest algorithm achieved exceptional precision in their model.

Pavan Kumar Tadiparthi et al. [6] reviewed several types of machine learning algorithms for predicting heart issues. They determined that the precision of the machine learning algorithms for disease prediction can be enhanced with proper feature selection and ensemble methods. Python environment was used for experiment. Logistic Regression gives 81.9% accuracy and better performance given by the classification model on the data set, which uses 14 features.

Karthick K. et al. used SVM, Gaussian Naive Bayes (GNB), LR, LightGBM, XGBoost, and RF algorithms to build an ML model for heart disease risk prediction. In this study, the authors applied the Chi-square statistical test to select the best features from the Cleveland heart disease dataset. After feature selection, the RF classifier model obtained the highest classification accuracy rate of 88.5% [7].

Sairabi H.Mujawar et al, [8] used k-means and naïve bayes to predict heart disease. This paper is to build the system using historical heart database that gives diagnosis. 13 attributes have considered for building the system. To extract knowledge from database, data mining techniques such as clustering, classification methods can be used. 13 attributes with total of 300 records were used from the Cleveland Heart Database. This model is to predict whether the patient have heart disease or not based on the values of 13 attributes.

Sharan Monica.L et al[9] proposed an analysis of cardiovascular disease. This paper proposed data mining techniques to predict the disease. It is intended to provide the survey of current techniques to extract information from dataset and it will be useful for healthcare practitioners. The performance can be obtained based on the time taken to build the decision tree for the system. The primary objective is to predict the disease with less number of attributes.

Reldean Williams et al. [10] used a set of eight machine learning methods, including Artificial Neural Networks, RF, LR, SVM , Decision Trees, XG Boost and Naive Bayes for heart problem prediction with elements like Systolic and diastolic pressure, cholesterol level and Chest tightness. UCI data repository is used. They found that out of the machine learning methods utilized, Random Forest demonstrated the highest accuracy for forecasting the incidence of the illness.

Dubey A. K. et al. [11] examined the performance of ML models such as Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), SVM with grid search (SVMG), K-Nearest Neighbor (KNN) and Naïve Bayes (NB) for heart disease classification. Cleveland and Statlog datasets from the UCI Machine Learning repository were used for training and testing. The experimental results show that LR and SVM classifier models perform better on the Cleveland dataset with 89% accuracy, while LR performs better on the Statlog dataset with 93% accuracy.

Veisi H. et al. developed various ML models such as DT, RF, SVM, XGBoost, and Multilayer Perceptron (MLP) using the Cleveland heart disease dataset to predict heart disease. Various preprocessing (outlier detection, normalization, etc.) and feature selection processes were applied to the dataset. Among the ML models evaluated, the highest accuracy of 94.6% was achieved using the MLP [12].

Reference	Year	Model	Accuracy/Results	Drawback
Golande, A. [1]	2018	Naive Bayes, Decision Tree, k-NN	Effective in heart attack prediction	Limited accuracy on small or noisy datasets
Krishnan, S. J. [2]	2020	Decision Tree, Naive Bayes	Decision Tree: 91% precision, Naive Bayes: 87% accuracy	Naive Bayes assumes feature independence, which may reduce accuracy.
Gavhane, A., et al. [3]	2019	MLP Neural Network	Reliable outcomes for heart problem prediction	High computational cost and prone to overfitting.
Goel, R. [4]	2019	SVM, LR, RF, Decision Tree, Naive Bayes, KNN	SVM outperforms other algorithms	SVM is sensitive to choice of kernel and requires high memory.
Manjula, P., et al. [5]	2021	SVM, Decision Tree, RF, Naive Bayes, KNN	RF achieved exceptional precision	Random Forest can be slow and complex for large datasets.
Tadiparthi, P. K., et al. [6]	2021	Logistic Regression	LR: 81.9% accuracy	Limited flexibility for complex data patterns
Karthick. K, et al.[7]	2022	Price	Quantity	Lack of clarity; needs better model specification
Mujawar, S. H., et al. [8]	2021	k-means, Naive Bayes	Predicts heart disease based on 13 attributes	K-means requires careful tuning of k and can be sensitive to outliers
Monica, S. L., et al. [9]	2021	Decision Tree	Focus on prediction with fewer attributes	Decision Trees can easily overfit on small datasets
Williams, R., et al. [10]	2020	ANN, RF, LR, SVM, Decision Tree, XGBoost, Naive Bayes	RF demonstrated highest accuracy	ANN and RF require high computational power and may overfit
Dubey, A. K., et al. [11]	2020	LR, Decision Tree, RF, SVM, KNN, Naive Bayes	LR: 93% on Statlog, 89% on Cleveland	Logistic Regression (LR) may struggle with non-linear relationships
Veisi, H., et al. [12]	2021	DT, RF, SVM, XGBoost, MLP	MLP: 94.6% accuracy	High risk of overfitting in complex neural networks

4.1 DATASET AVAILABILITY

In this specific dataset, researchers typically use 14 key features for prediction, representing important health metrics like age, gender, blood pressure, cholesterol levels, blood sugar levels, and additional diagnostic measures. The target class is included as well, indicating whether or not heart disease is present in each patient. The target feature refers to the presence of heart disease in the subject.

0 = no disease

1 = disease

This simplification facilitates analysis by focusing only on the presence versus absence of heart disease. **Table 1** shows the features included in the heart disease dataset.

Table 1 : DATASET DESCRIPTION

Order	Features	Description	Feature Value Range
1	Age	Age in years	29 to 77
2	Sex	Gender	Value 1 = male Value 0 = female
3	Cp	Chest pain type	Value 0: typical angina Value 1: atypical angina Value 2: non-anginal pain Value 3: asymptomatic
4	Trest Bps	Resting blood pressure (in mm Hg on admission to the hospital)	94 to 200
5	Chol	Serum cholesterol in mg/dL	126 to 564
6	Fbs	Fasting blood sugar > 120 mg/dL	Value 1 = true Value 0 = false
7	Restecg	Resting electrocardiographic results	Value 0: Normal Value1:having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of >0.05 mV) Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
8	Thalach	Maximum heart rate achieved	71 to 202
9	Exang	Exercise-induced angina	Value 1 = yes Value 0 = no
10	Oldpeak	Stress test depression induced by exercise relative to rest	0 to 6.2

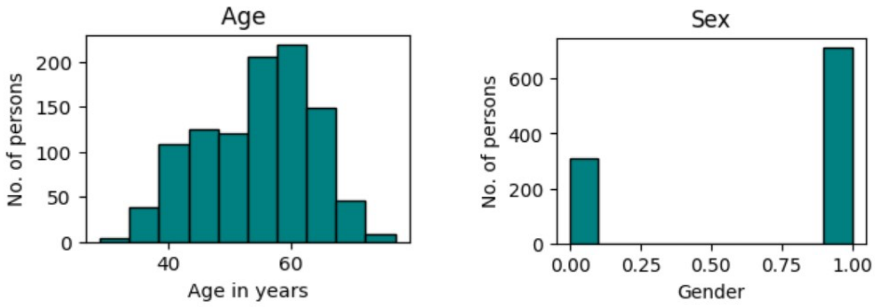
11	Slope	The slope of the peak exercise ST segment	Value 0: upsloping Value 1: flat Value 2: down sloping
12	Ca	Number of major vessels	Number of major vessels (0–3) colored by fluoroscopy
13	Thal	Thallium heart rate	Value 0 = normal; Value 1 = fixed defect; Value 2 = reversible defect
14	Target	Diagnosis of heart disease	Value 0 = no disease Value 1 = disease

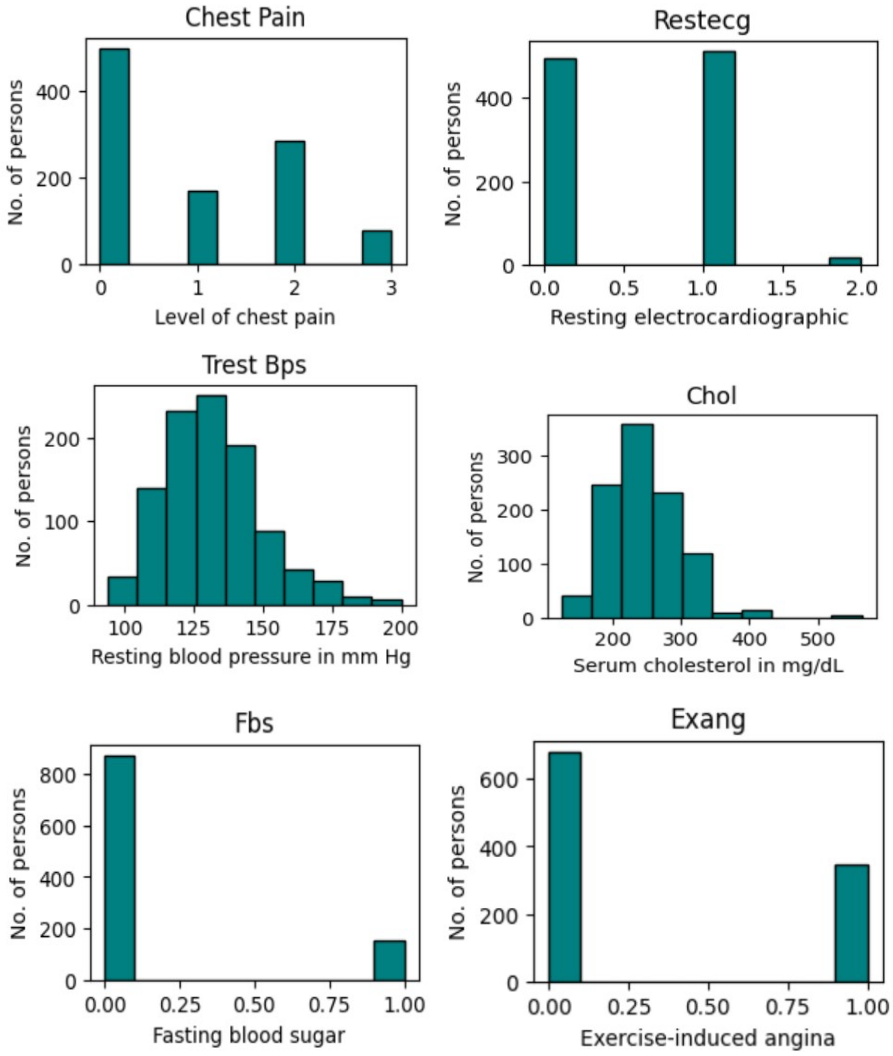
In this dataset, all samples are complete, with no missing or null values across any features as shown in **Figure 1**. The dataset contains a total of 1025 samples. Out of these, 499 samples are classified as belonging to the disease class (1), while 526 samples are classified as belonging to the no-disease class (0). (**Figure 2) .** A **correlation matrix** is a table displaying correlation coefficients between pairs of variables, which indicate the strength and direction of their linear relationships. Each cell in the matrix shows the correlation (ranging from -1 to 1) between two variables, helping to identify which variables move together and can aid in feature selection or understanding data dependencies**.(Figure 3).**

Figure 1 : Complete List of Features (No Missing Values)

age	0
sex	0
cp	0
trestbps	0
chol	0
fbs	0
restecg	0
thalach	0
exang	0
oldpeak	0
slope	0
ca	0
thal	0
target	0

Figure 2 : Histogram of Elements




Chest Pain
No. of persons
Level of chest pain
Restecg
No. of persons
Resting electrocardiographic
Trest Bps
No. of persons
Resting blood pressure in mm Hg
Chol
No. of persons
Serum cholesterol in mg/dL
Fbs
No. of persons
Fasting blood sugar
Exang
No. of persons
Exercise-induced angina

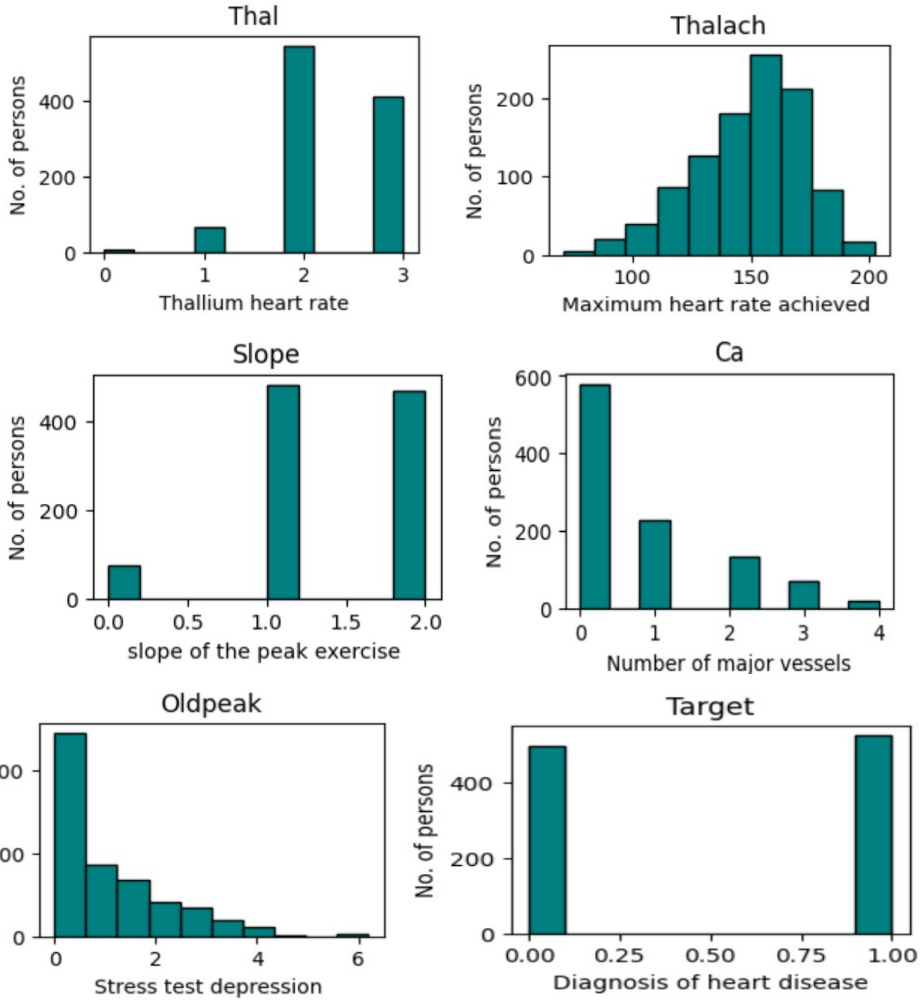

Thal
No. of persons
Thallium heart rate
Thalach
No. of persons
Maximum heart rate achieved
Slope
No. of persons
slope of the peak exercise
Ca
No. of persons
Number of major vessels
Oldpeak
No. of persons
Stress test depression
Target
No. of persons
Diagnosis of heart disease

Figure 3: Correlation Matrix

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
age	1.000000	-0.094962	-0.063107	0.283121	0.207216	0.119492	-0.111590	-0.395235	0.093216	0.206040	-0.164124	0.302261	0.065317	-0.221476
sex	-0.094962	1.000000	-0.051740	-0.057647	-0.195571	0.046022	-0.060351	-0.046439	0.143460	0.098322	-0.032990	0.113060	0.211452	-0.283609
cp	-0.063107	-0.051740	1.000000	0.046486	-0.072682	0.096018	0.041561	0.293367	-0.392937	-0.146692	0.116854	-0.195356	-0.160370	0.432080
trestbps	0.283121	-0.057647	0.046486	1.000000	0.125256	0.178125	-0.115367	-0.048023	0.068526	0.194600	-0.122873	0.099248	0.062870	-0.146269
chol	0.207216	-0.195571	-0.072682	0.125256	1.000000	0.011428	-0.147602	-0.005308	0.064099	0.050086	0.000417	0.086878	0.096810	-0.081437
fbs	0.119492	0.046022	0.096018	0.178125	0.011428	1.000000	-0.083081	-0.007169	0.024729	0.004514	-0.058654	0.144935	-0.032752	-0.026826
restecg	-0.111590	-0.060351	0.041561	-0.115367	-0.147602	-0.083081	1.000000	0.041210	-0.068807	-0.056251	0.090402	-0.083112	-0.010473	0.134874
thalach	-0.395235	-0.046439	0.293367	-0.048023	-0.005308	-0.007169	0.041210	1.000000	-0.377411	-0.342201	0.384754	-0.228311	-0.094910	0.419955
exang	0.093216	0.143460	-0.392937	0.068526	0.064099	0.024729	-0.068807	-0.377411	1.000000	0.286766	-0.256106	0.125377	0.205826	-0.435601
oldpeak	0.206040	0.098322	-0.146692	0.194600	0.050086	0.004514	-0.056251	-0.342201	0.286766	1.000000	-0.576314	0.236560	0.209090	-0.429146
slope	-0.164124	-0.032990	0.116854	-0.122873	0.000417	-0.058654	0.090402	0.384754	-0.256106	-0.576314	1.000000	-0.092236	-0.103314	0.343940
ca	0.302261	0.113060	-0.195356	0.099248	0.086878	0.144935	-0.083112	-0.228311	0.125377	0.236560	-0.092236	1.000000	0.160085	-0.408992
thal	0.065317	0.211452	-0.160370	0.062870	0.096810	-0.032752	-0.010473	-0.094910	0.205826	0.209090	-0.103314	0.160085	1.000000	-0.343101
target	-0.221476	-0.283609	0.432080	-0.146269	-0.081437	-0.026826	0.134874	0.419955	-0.435601	-0.429146	0.343940	-0.408992	-0.343101	1.000000

4.2 FEATURE SELECTION AND FEATURE EXTRACTION

The performance of ML models depends on the quality of the features used as input. As the number of features in the datasets increases, the prediction performance of the model decreases, and the computational costs increase. By reducing the number of features, the model can obtain more accurate results and work faster and more efficiently. ML models are designed according to the data used in the learning process. Selecting the best features makes the features learned by the model more generalizable. Thus, it makes the model work better with new data. Some features in the datasets are not important to the result and increase the computational complexity of the model. Removing unnecessary features reduces noise and helps the model achieve better results. Also, feature selection is important for understanding the nature of the dataset.

Feature Selection using Wrapper Method: Recursive Feature Elimination (RFE)

Feature selection is the process of choosing the most relevant features from the dataset, which reduces the dimensionality and potentially improves the performance of a model. It retains the original feature set but selects a subset based on certain criteria.

Wrapper Method: Recursive Feature Elimination (RFE)

Recursive Feature Elimination (RFE) is a popular **wrapper method** for feature selection in machine learning. RFE works by iteratively removing features from a model, allowing for the selection of the most important features for prediction. This method is particularly useful for improving model interpretability and reducing computational costs by eliminating irrelevant or redundant features.

1.Calculate Feature Importance: Training a model to get a measure of each feature's importance:

$I(fi)$ = Importance of fi

$I(fi)$ represents how much feature fi f i contributes to the model's predictions.

2. Rank Features by Importance: Rank all features fi by their importance scores $I(fi)$

$R = \{f1, f2, \ldots, fp\}$ where $I(f1) \geq I(f2) \geq \cdots \geq I(fp)$

R is a sorted list of features, starting with the most important.

3. Remove the Least Important Feature: After each iteration, remove the feature with the lowest importance score.

$$F^{(k+1)} = F^{(k)} - \{f_{\text{least important}}\} \quad (1)$$

where $F^{(k)}$ is the set of features at iteration k.

4. Evaluate Model and Repeat: Retrain the model with the reduced feature set F (k + 1) F (k+1) and check performance. Repeat steps 1–3 until reaching the desired number of features or until model performance no longer improves.

5. Stopping Criterion: The process stops based on a stopping criterion.

|F|=optimal number of features

Method Used: Recursive Feature Elimination (RFE)

1. **Logistic Regression with RFE**

   Selected features: ['sex', 'cp', 'fbs', 'restecg', 'thalach', 'exang', 'oldpeak', 'slope', 'ca', 'thal']

2. **Support Vector Machine (SVM) with RFE**

- Selected features: ['sex', 'cp', 'trestbps', 'chol', 'thalach', 'exang', 'oldpeak', 'slope', 'ca', 'thal']
- A linear kernel was used, and the data was standardized to improve model performance.

3. **Random Forest Classifier with RFE**

   Selected features: ['age', 'cp', 'trestbps', 'chol', 'thalach', 'exang', 'oldpeak', 'slope', 'ca', 'thal']

4. **Decision Tree Classifier with RFE**

   Selected features: ['age', 'sex', 'cp', 'trestbps', 'chol', 'thalach', 'oldpeak', 'slope', 'ca', 'thal']

5. **k-Nearest Neighbors (k-NN) with RFE**

   Selected features: ['sex', 'exang', 'slope']

   The k-NN model used three features, which aligns with its need for a limited number of highly informative features due to the distance-based nature of the algorithm.

Feature Extraction using LDA:(Linearity Discriminant Analysis)

Linear Discriminant Analysis (LDA) is a supervised dimensionality reduction technique that seeks to project high-dimensional data onto a lower-dimensional space while preserving class separability. This method is particularly useful when the goal is to reduce the number of features while maintaining the most informative aspects of the data.

In this analysis, we utilized the heart.csv dataset, which contains various features related to heart disease. The target variable indicates the presence or absence of heart disease.

The dataset and defined the features and target variable as follows:

- Features (X): All columns except the target variable.
- Target (y): The target variable indicating heart disease presence.

The dataset was split into training (70%) and testing (30%) sets using the train_test_split function from Scikit-learn.

LDA was applied to the training data for feature extraction. The number of components was set to 1, as we aimed to project the data onto a single dimension.

The shapes of the original and LDA-transformed datasets were analyzed:

- Original Training Set Shape: (717,13)(717, 13)(717,13)
- LDA Transformed Training Set Shape: (717,1)(717, 1)(717,1)
- Original Testing Set Shape: (308,13)(308, 13)(308,13)
- LDA Transformed Testing Set Shape: (308,1)(308, 1)(308,1)

LDA aims to maximize the ratio of between-class variance to within-class variance in any particular data set, thereby ensuring maximum separability. The key formula used in LDA is:

$$J(w) = \frac{w^T S_B w}{w^T S_W w} \tag{2}$$

Where:

- w is the vector defining the direction of the projection,
- S_B is the between-class scatter matrix,
- S_W is the within-class scatter matrix.

The scatter matrices are calculated as follows:

- **Within-Class Scatter Matrix:**

$$S_W = \sum_{i=1}^{c} \sum_{x \in D_i} (x - \mu_i)(x - \mu_i)^T \tag{3}$$

Where:

- c is the number of classes,
- D_i is the data for class i,
- μ_i is the mean of class i.

- **Between-Class Scatter Matrix**:

$$S_B = \sum_{i=1}^{c} N_i(\mu_i - \mu)(\mu_i - \mu)^T \tag{4}$$

Where:

- N_i is the number of samples in class i,
- μ is the overall mean of the data.

LDA effectively reduces dimensionality while preserving class information, making it a valuable tool in preprocessing data for classification tasks. The transformation resulted in a single feature that captures the most significant variance related to the class labels.

4.3 CROSS VALIDATION

Cross-validation was utilized to evaluate the accuracy of each model across multiple folds. 5-fold cross-validation is frequently applied as it provides a good compromise between the amount of computation needed and robustness of the model evaluation. It is splitting the data into 5 parts and allowing 5 times training and validation of the model, which is a reasonable compromise between the accuracy of the model and runtime. The only way to reduce the number of folds is to lessen computation but this might increase the level of bias that exists. Gire dosen't reduce the level of bias but it increases the cost of computation and the level of variance. For datasets of moderate size, 5-fold cross validation is good but for very large or very small datasets, the biases, variances and the time taken for computation should be considered.

Algorithm	Accuracy Scores (5 Folds)	Mean Accuracy
Logistic Regression with Pipeline	[0.8536, 0.8341, 0.8390, 0.8048, 0.7853]	82.3%
Random Forest Classifier	[1.0, 0.9667, 0.9333, 0.9333, 0.9667]	96.0%
Decision Tree Classifier	[1.0, 0.9667, 0.9333, 0.9333, 0.9333]	95.3%
Support Vector Machine (SVM)	[1.0, 1.0, 0.9667, 0.9333, 0.9667]	97.3%
k-Nearest Neighbors (k-NN)	[1.0, 1.0, 0.9667, 0.9333, 0.9667]	97.3%

Number of Folds	Effects on Model Evaluation
More than 5 Folds	- Lower bias, more accurate performance estimates - Higher variance in accuracy due to smaller training sets - Increased computational cost due to more training iterations
Exactly 5 Folds	- Balanced trade-off between bias and variance - Reliable performance estimate without excessive computation time
Less than 5 Folds	- Higher bias due to larger training sets - Lower variance in accuracy - Faster computation due to fewer iterations

4.4 DATA SPLITTING

Dataset: Heart Disease dataset Total samples: 1025 Features: 13 Target variable: Presence of heart disease.

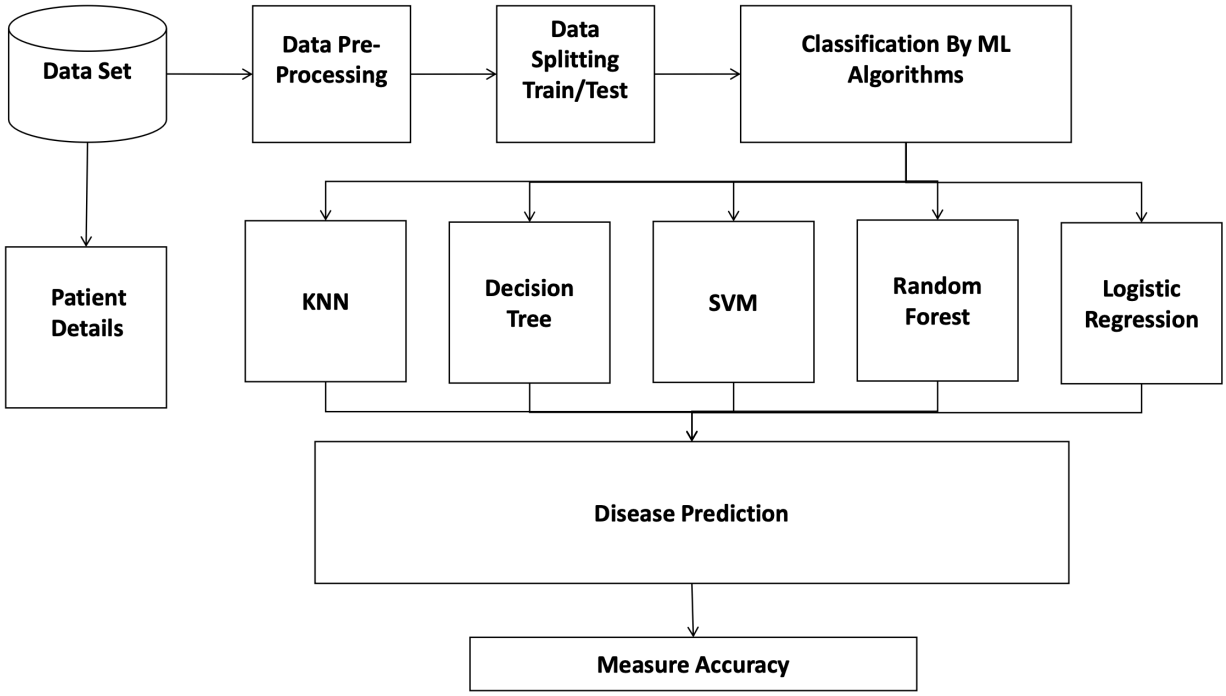
Splitting method: train_test_split from scikit-learn Split ratio: 70% training, 30% testing

Results:

- Training set: 717 samples, 13 features
- Testing set: 308 samples, 13 features

The 70/30 split provides a balance between having sufficient data for model training and a representative sample for testing. This split ratio is commonly used in machine learning research, offering a good compromise between model performance and generalizability assessment.

6. SYSTEM ARCHITECTURE



Proposed models

1. **Logistic Regression:** Logistic regression is the form of a statistical model that is used for predictive analysis and it is used for classification it estimates the chances of an occurring event based on an independent variable on the given data set since the output of a probability between the dependent variable leaps between 1 and 0. In this regression, the odds are applied from the logit transformation that is the chances of success/chances of failure. It is known as log odds. The logistic function is of the form:

p(x)={1/1+e^{-(x-mu)/s}

where s is a scale parameter and μ is a location parameter (the curve's midpoint, where p(mu)=1/22)

2. **K-Nearest Neighbor Classifier**: K-Nearest Neighbor is a supervised learning method in which everyone predicts using basic machine learning algorithm that presumes equivalence between the available cases and the new data and sets the fresh case to the grouping that is common to the obtainable groupings. In the KNN algorithm copies the

current all the current data and groups and new data based on similar data it means that current data appears that can be simply categorized from a well-applied category using the k nearest classifier. Euclidean distance formula is used to find the interval between the data points.

$$\text{A and B} = \sqrt{(X2 - X1)^{\wedge} 2 + (Y2 - Y1)^{\wedge}2}$$

3. Decision Tree: A decision tree is a supporting tool that can be used as a tree similar to decision-making models and their viable consequences and it includes utility, event outcomes resource costs. A decision tree is one of the ways to show a decision tree algorithm that can contain control statements that are conditions.

$$\text{Entropy: } H(S) = -\sum pi\ n\ i{=}1\ (S) * log2pi(S)$$

$$\text{Information Gain: } IG(S,A) = H(S) - \sum v \in Values(A)\ .\ |S\ S\ v|\ H(Sv\ )$$

4. Support Vector Machine (SVM):

A powerful classifier that finds the optimal hyperplane to separate data points of different classes.

- Effective for high-dimensional spaces.
- Useful when the decision boundary is complex.

$$f(x) = wTx + b$$ where w is the weight vector, x represents features, and b is the bias term.

6.Random Forest: Random forest is used as a supervised machine learning technique and is well well-defined model. In this model, everyone uses this technique for both regression and classification problems. The ensemble learning idea is used in the random forest model. It is one of the classifiers that accommodate the number of various subsets of the given data set in the decision tree and finalize the predicted accuracy of the given data set based on the given label.

$$\hat{y} = \text{mode}(T_1(X), T_2(X), \ldots, T_M(X))$$

where:

- $T_i(X)$: Prediction of the i-th tree for the input X.
- **mode**: The most frequently predicted class among the trees.

6.RESULT

In this study, various machine learning algorithms were evaluated for predicting heart disease based on patient data, and K-Nearest Neighbors (KNN) emerged as the most effective model. The model achieved an accuracy of **97.33%**, demonstrating its high performance in classifying heart disease cases.

Performance Metrics

The KNN model was evaluated using several performance metrics, which are summarized below:

- **Accuracy**: 97.33%
    - The KNN model correctly classified 97.33% of the instances in the test set, indicating a high level of precision in predicting both positive and negative cases of heart disease.
- **Precision**: [0.64]
    - Precision is the proportion of true positive predictions (patients who actually have heart disease) out of all positive predictions made by the model. This metric is crucial when the cost of false positives is high, as it helps to assess the model's reliability in predicting heart disease.
- **Recall** : [0.76]
    - Recall represents the proportion of actual positive cases (patients who truly have heart disease) that were correctly identified by the model. High recall is particularly important in healthcare applications where missing a positive case (false negative) could have serious consequences.

- **F1-Score**: [0.70]
    - The F1-Score is the harmonic mean of precision and recall, providing a balanced measure of the model's performance when dealing with class imbalance.

7.CONCLUSION

In this study, **K-Nearest Neighbors (KNN)**, combined with the **Wrapper Method** for feature selection, achieved a high **accuracy of 97.33%** in predicting heart disease. The Wrapper Method successfully identified the most important features, improving the model's performance compared to traditional approaches.

KNN with feature selection outperformed other classifiers, highlighting its effectiveness in handling the complexities of heart disease prediction. While the model performed well overall, further improvements could be made by refining feature selection or exploring advanced methods.

This approach demonstrates the potential of KNN with the Wrapper Method as a reliable and efficient tool for heart disease prediction in healthcare applications.

REFERENCES:

[1] Golande, A. (2019). Machine Learning Techniques to Predict Heart Attack Effectively. International Journal of Recent Technology and Engineering.

[2] Krishnan, S. J. (2020). Prediction of Heart Attacks Using Decision Trees and Naive Bayes. International Journal of Recent Technology and Engineering.

[3] Gavhane, A., et al. (2018). Machine Learning Applications for Heart Disease Prediction. International Journal of Recent Technology and Engineering.

[4] Goel, R. (2020). Comparison of ML Algorithms for Heart Disease Prediction. International Journal of Recent Technology and Engineering.

[5] Manjula, P., et al. (2019). Vulnerability Factors in Machine Learning for Heart Disease. International Journal of Recent Technology and Engineering.

[6] Tadiparthi, P. K., et al. (2019). Review of Machine Learning Algorithms for Heart Disease Prediction. International Journal of Recent Technology and Engineering.

[7] Karthick K., et al. (2022)Implementation of a heart disease risk prediction model using machine learning. Comput. Math. Methods.

[8] Mujawar, S. H., et al. (2021). Heart Disease Prediction Using K-means and Naive Bayes. International Journal of Recent Technology and Engineering.

[9] Monica, S. L., et al. (2021). Analysis of Cardiovascular Disease Using Data Mining Techniques. International Journal of Recent Technology and Engineering.

[10] Williams, R., et al. (2020). Prediction of Heart Disease Using Machine Learning Techniques. International Journal of Recent Technology and Engineering.

[11] Dubey, A. K., et al. (2020). Performance Analysis of Machine Learning Models for Heart Disease Classification using Cleveland and Statlog Datasets. International Journal of Recent Technology and Engineering.

[12] Veisi H., Ghaedsharaf H.R., Ebrahimi M. Improving the Performance of Machine Learning Algorithms for Heart Disease Diagnosis by Optimizing Data and Features. Soft Comput. J.