

HEART DISEASE PREDICTION SYSTEM USING MACHINE LEARNING ALGORITHMS

*Muhammad Abrar*¹, *Muhammad Sadique*²

1 Department of Computer Science GPGC Mansehra, Pakistan

2 Department of Computer Science GPGC Mansehra, Pakistan

ABSTRACT

One of the most difficult problems in the medical field nowadays is the prediction of heart disease. Approximately in the current period, one person dies from heart disease every minute. In the sphere of healthcare, the process must be automated to reduce risks and notify the patient well in advance because predicting cardiac illness is a challenging task. In this study, the UCI machine learning repository's heart disease data-set is used.

In the current project, five well-known machine learning algorithms—decision tree, Random forest, Support Vector Machine, logistic regression, and k-neighbor—are being evaluated for accuracy by using machine learning, which learns from natural events and data and uses biological variables like cholesterol, blood pressure, sex, and age as evaluating data. Finding the most accurate way to predict cardiac illnesses is the study's main objective.

The performance of five well-known machine learning algorithms is thus analyzed in this work to give a comparison study. In comparison to other ML algorithms used, the testing results demonstrate that the support vector machine approach has the highest accuracy of 85%.

Keywords— *Decision tree, Support vector machine, linear regression, decision tree, naïve bayes, supervised, unsupervised, reinforced,*

1. INTRODUCTION:

The heart is a crucial part of the human body that propels blood through the vessels of the cardiovascular system. The circulatory system's primary purpose is to transport essential elements such as blood, oxygen, and other substances to different organs throughout the body. As a result, correct cardiac function is critical, as poor performance can lead to serious health issues and even death.

In Pakistan, Cancer has been eclipsed by cardiovascular illness as the leading cause of death, with heart attacks increasingly impacting people under the age of 30. This increase in

deaths among individuals aged 30 to 40 can be attributed to a variety of risk factors, including the country's widespread sedentary lifestyle and frequent smoking. Notably, young adults in this age bracket are more prone to hypertension, a condition that can cause a heart attack. Heart attacks among young people have increased in Pakistan as a result of rising tobacco usage and smoking rates [1].

To combat the growing epidemic of heart attacks in Pakistan, a serious effort must be made to address key risk factors such as hypertension, diabetes, smoking, and high cholesterol levels. Steps must be taken to reduce the country's rising incidence of cardiovascular disease.

According to data from the World Health Organization (WHO) for 2016, cardiac disease is responsible for 19% of all fatalities in Pakistan, resulting in approximately 250,000 deaths each year. The WHO has observed a significant 29% increase in Pakistan's overall mortality rate related to heart disease, leading to an alarming figure of nearly 406,870 deaths per year in just three years [1].

Machine learning, a subfield of artificial intelligence (AI) that permits computers to mimic human abilities, is a powerful instrument that can aid in the accurate diagnosis and management of cardiac illness. Machine learning systems can be taught to analyze and comprehend data through training and testing, allowing for more effective and accurate detection and management of cardiac issues. The combination of AI and machine learning, known as machine intelligence, offers tremendous potential for the detection and prevention of cardiac diseases all over the world.

In the current project, the five most famous algorithms are being assessed for accuracy using machine learning, which learns from patient data and uses biological variables like cholesterol, blood pressure, sex, and age as evaluating data. The algorithms are decision tree, random forest, support vector machine, linear regression, and k-neighbor. The major goal of the study is to identify the most reliable method of cardiac disease prediction.

We have structured our paper in the following steps;

- In Section-(I) of the study machine learning and its application to heart diseases are explained.
- In Section-(II) deals with classification machine learning method used in our research.
- In Section-(III) already explained researcher's work is elaborated.
- In Section-(IV) our proposed model is discussed
- In Section-(V) we describe the models used in our project.
- In Section-(VI) we discussed the data-set we have used and how our project significant it is according to our dataset

- In Section-(VII) the study ends with a summary that also discusses the project's possible future applications.

Overall, the initiative seeks to use machine learning to create more accurate predictive models for cardiac diseases, with the potential to better patient detection and therapy results.

2. MACHINE LEARNING

Modern technology known as "machine learning" makes use of testing and training to improve a computer system's success in a particular disease prediction. The system gains knowledge and enhances its capabilities by receiving instruction using data and experiences. While the training phase includes the system learning from a training data-set to better its performance on the job at hand, the testing phase is crucial for determining the system's accuracy and generalization ability when applied to a new dataset

The main categories of machine learning methods are reinforcement learning, unsupervised learning, and supervised learning. Each kind has a set of techniques for dealing with certain types of disease to predict decisions. Now we will explain each type in detail as shown in Figure 1 [2].

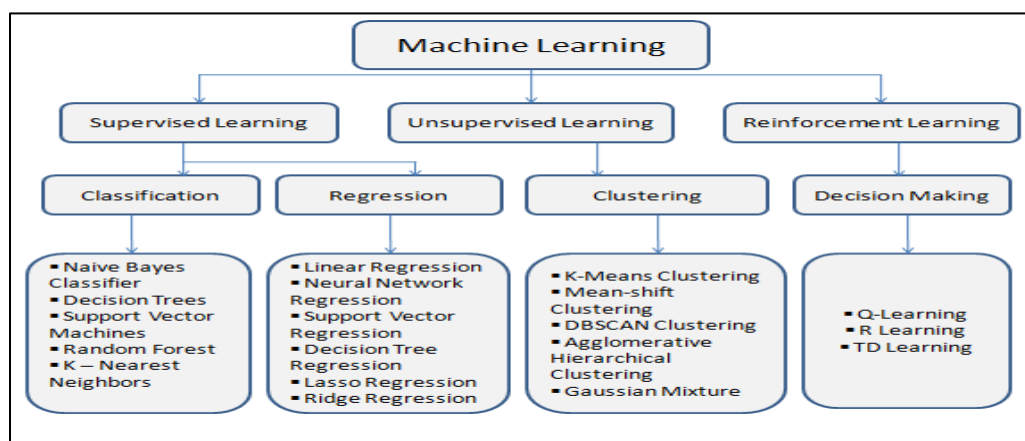


Figure 1 Machine learning types

A. Supervised Learning

Through supervised learning, a computer is taught to learn using an annotated dataset. In this process, a training data-set acts as a manual or tutor for the system as it makes predictions on fresh data while being tested by supervised learning algorithms, which are based on the "teach me" concept, employ techniques including classification, regression, decision trees, and random forests.

Regression analysis is used to identify patterns and determine the probability of continuous events. This includes the system's ability to identify numerical numbers and integrate numerical senses like width and height.

The most widely used guided machine learning methods are Naive Bayes, Decision Trees, Neural Networks, Support Vector Machines (SVM), Random Forests, Logistic Regression, and Logistic Regression with Gradient Boosted Trees. These equations are applied to a variety of problems in industries including finance, healthcare, transportation, and others.

B. Unsupervised Learning

Unsupervised learning is a sub-type of machine learning in which the algorithm is not provided with instructions or training data that has been annotated. The application utilizes the dataset to automatically analyze it for patterns and relationships between the data items. These links are employed to classify fresh data into one of the existing clusters. The idea of "self-sufficiency," in which the system identifies patterns on its own, is the foundation of unsupervised learning. For instance, we may divide a dataset of different fruits into three categories using unsupervised learning based on their similarities and differences, such as mango, banana, and apple. The system will then promptly send any new data to the relevant node.

Unsupervised learning techniques make use of grouping and dimensionality reduction procedures. By applying dimensionality reduction, the collection may retain all of the relevant data while having fewer features. Comparable data pieces are clustered together during the clustering process based on their similarities and differences.

Unstructured machine learning techniques that are often utilized include T-SNE, k-means clustering, and PCA. These techniques are used by many recognitions. Unsupervised learning is a powerful tool for machine learning that may be used to uncover patterns and relationships in data. Applications for anomaly detection, suggestion systems, image, and voice.

C. Reinforcement

Reinforcement learning is a machine learning technique where an agent interacts with its environment to learn how to make the best decisions. In contrast to supervised learning, there is no teacher and no annotated collection. Instead, the agent learns by making errors and then correcting them, gaining rewards or incurring penalties as a result of its actions in the outside world. The goal is to maximize the cumulative reward over a long period via reinforcement learning.

The agent gets knowledge by modifying its policy, which is a set of rules that directs the activities the agent does in a particular circumstance. The reinforcement learning process involves the agent, the external environment, and the reward indicator. The agent makes decisions based on its current policy while receiving input from the outside environment in the form of rewards or penalties. The agent adapts its approach in light of this feedback to make better decisions in the future.

3. RELATED WORK:

The heart, a vital component of the human body, circulates blood throughout the circulatory system. There is a lot of focus on safeguarding the heart from illness and malfunction because of its crucial part in maintaining life. As a result, many scholars are working on this project. An important component of this study is the examination of heart-related data, including diagnosis and prediction.

Artificial intelligence, machine learning, and deep learning have all made significant contributions to this research by developing novel techniques and algorithms for the analysis of heart-related data. Now we will discuss the most renowned work take on the same problem.

- i. In their research, [3] Rushottam et al. applied decision trees and hill-climbing algorithms. They used the Cleveland data-set to prepare the data for classification techniques. The knowledge extraction is carried out using the open-source data mining application KEEL, which completes the gaps in the data gathering. Decision trees use an up-down order. For each real node chosen by the hill-climbing algorithm at each stage, a test chooses a node. The factors are confidence and their associated numbers. It is at least 0.25 percent confident. About 86.7% of the time, the technique is accurate.
- ii. The [4] is another article that they recommend reading. In their research, they explain Naive Bayes and decision tree classifiers in detail and show how they are used specifically to predict heart disease. According to some studies that looked at the use of a predictive

- data mining method on the same data-set, Decision Trees have better accuracy than Bayesian models.
- iii. Senthil Kumar Mohan and associates published "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques" to increase the precision of heart problems. A heart disease forecasting model that incorporates a random forest and a linear model was created using the techniques KNN, LR, SVM, and NN. The result is an improved performance level with a degree of precision of 88.7%. (HRFLM). [5]
 - iv. In this study, S. Chellammal and R. Sharmila aimed to identify the most important characteristics for forecasting heart illness using correlation measures. They looked at 303 individuals in a data-set that contained 14 characteristics, including age, sex, blood pressure, cholesterol, the sort of chest discomfort, maximum heart rate, and others. The strength of the connection between each trait and the objective variable (presence of heart disease) was assessed by the writers using the Pearson correlation coefficient. The authors' research revealed that the factors most significantly associated with heart disease were age, sex, the sort of chest pain, the highest heart rate attained, and exercise-induced angina. Using these five characteristics, they were able to forecast heart illness with an accuracy of 85%. The efficacy of the suggested set of characteristics was also evaluated in comparison to other widely employed models, like support vector machines and decision trees, and the authors discovered that their model beat the others. Overall, the study highlights the importance of selecting appropriate attributes for predictive modeling in healthcare and provides a useful set of attributes for accurate prediction of heart disease. [3]
 - v. In their paper, the authors examine the machine learning application methods for the prediction of heart illness. The authors compare a variety of types, including decision trees, support vector machines, and k-nearest neighbors, to determine the most effective method. A support vector machine with an accuracy of 94.12% beats other models, according to the research. The outcomes show that machine learning systems have a chance to accurately forecast heart disease. [4]
 - vi. In their article, Umarani Nagavelli, Debabrata Samanta, and Partha Chakraborty describe how they used decision trees, logistic regression, and k-nearest neighbor algorithms to create machine learning-based models for detecting cardiac disease. Using these algorithms, the scientists predicted heart illness with an 87% accuracy using a data-set of individuals with different characteristics. The research demonstrates the utility of machine learning in healthcare and offers a practical method for precise heart disease diagnosis. [6]
 - vii. The authors in [7] examined the use of machine learning methods for cardiac illness prediction. The article explains how machine learning techniques can be used to forecast

- heart disease. The writers collected a group of patients with a variety of characteristics such as age, gender, blood pressure, cholesterol level, and so on, and developed several models such as k-nearest neighbor, logistic regression, decision tree, and random forest. Their study shows that the random forest method outperforms other models by 90%. The results imply that machine learning systems can provide accurate predictions of heart illness. According to the research, using machine learning in healthcare can significantly improve diagnosis precision and patient outcomes.
- viii. This article describes a machine learning-based technique for identifying heart disease in e-healthcare that employs classification algorithms to forecast the likelihood of heart disease based on patient data like gender, age, and blood pressure. The study collected patient data from Pakistani institutions and used various categorization algorithms, such as Naive Baye, decision trees, and k-Nearest Neighbor, to determine the most effective approach. The k-Nearest Neighbors algorithm was found to be the most accurate in forecasting cardiac illness, with an accuracy rate of 86.5%. The method suggested here can be used in e-healthcare to detect heart illness early and enhance patient outcomes. [8].
- ix. This review article investigates the prediction of heart disease using machine learning methods. The frequency and effects of heart illness are emphasized throughout the article, along with the importance of early detection and diagnosis. The authors discuss the various artificial neural networks, decision trees, and support vector machines that are used in machine learning to predict cardiac illness. They evaluate the effectiveness of these programs by comparing their performance. According to the study's findings, machine learning methods can reliably forecast cardiac illness and help doctors diagnose and treat patients. This article offers insightful analyses of the state of the art in machine learning-based cardiac disease prediction study. [9]

4. METHODOLOGY OF SYSTEM

Our heart disease detection research project starts with data-set pre-processing and selecting attributes from data for machine learning models. A well-known trained machine learning model is then used to prediction of heart disease. Based on the prediction the outcomes are clearly shown for easier comprehension. Our project methodology comprises of following six steps as shown in Figure 2.

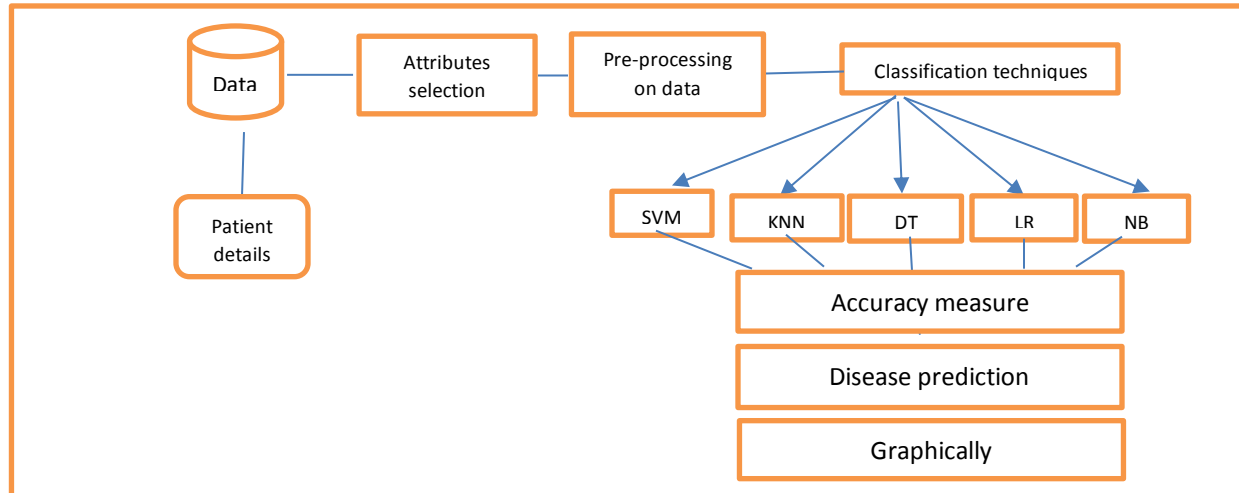


Figure 2 Architecture of prediction system

1) *Dataset Collection:*

We commence by collecting data for our cardiac illness prediction system. We separated the data into training and assessment after it had been gathered. Prediction models are constructed using both the assessment and training databases, and then they are assessed. 70% of the materials are used for instruction in this endeavor, with 30% being used for evaluation. In this investigation, the Heart Disease Kaggle (UCI (University of California, Irvine) data-set was used. Of the 303 patient data-set, the algorithm uses 14 of them.[10]

2) *Selection of attributes Attribute or Feature*

A stage in the choosing procedure is picking the appropriate predictor system characteristics. The efficacy of the method is increased by doing this. The forecast is based on several patient factors, such as gender, the sort of chest discomfort, basal blood pressure, serum cholesterol, exang, etc. The link matrix is used to select the characteristics of this model.

Details of attributes

S.no	Attributes	Value	Description
1	Age	29-62	Age in years
2	Sex	0-male,1-female	Gender
3	Cp	1 typical angina,2-atypical angina,3-non-anginal pain,4-asymptomatic	Chest pain type
4	restbps	Numerical value(140mm/hg)	Resting blood pressure in mm/hg
5	Chol	The numerical value (289mg/dl)	Serum cholesterol in mg/dl
6	Fbs	0-false,1-true	Fasting blood pressure >120mg/dl
7	restecg	1-normal,1-having-T,2-hypertrophy	Resting electrocardiographic results
8	thalach	140,173	Maximum heart rate achieved
9	exam	0-no,1-yes	Exercise-induced angina
10	Oldpeak	Numerical value	ST depression induced by exercise ST segment
11	Slope	1-upsloping,2-flat,3-downsloping	The slope of the peak exercise ST segment
12	Ca	0-3 vessels	Number of major vessels colored by fluoroscopy
13	Thal	3-normal,6-fixed defect,7-reversible defect	Thalassemia
14	Num	0 1	Diagnosis of heart disease(angiographic disease status)

Table 1 Details of attributes”

3) Pre-processing of Data

Pre-processing the data involves converting the data into the structure that we require, and dealing with the dataset's noise, duplication, and absent numbers, among other things. Pre-processing the data includes actions such as data-set input, data-set segmentation, attribute scaling, etc.

Data cleansing, data transformation, data integration, and feature reduction are the specific preprocessing techniques we use in our heart disease detection system.[17]

To function properly, the model needs the data to be accurate and, in the format, specified by the algorithm.

4.) *Balancing of Data*

Unbalanced datasets can be balanced using two methods. Out of 303 data, 203 of which were male and 100 of which were female, they are both under sampling and oversampling, as shown in Figure 5. Of the males, 110 have heart disease and 93 do not, while among the girls, 75 have heart disease and 25 do not.

There are two approaches to balance unbalanced knowledge. Both under- and over-sampling take place.

1. Under Sampling: Under Sampling reduces the size of the abundant class to attain data-set equilibrium. This process is taken into consideration when the data volume is adequate.
2. Over Sampling: Over Sampling increases the number of limited samples to bring data-set equilibrium. This process is looked at when the amount of info accessible is insufficient.

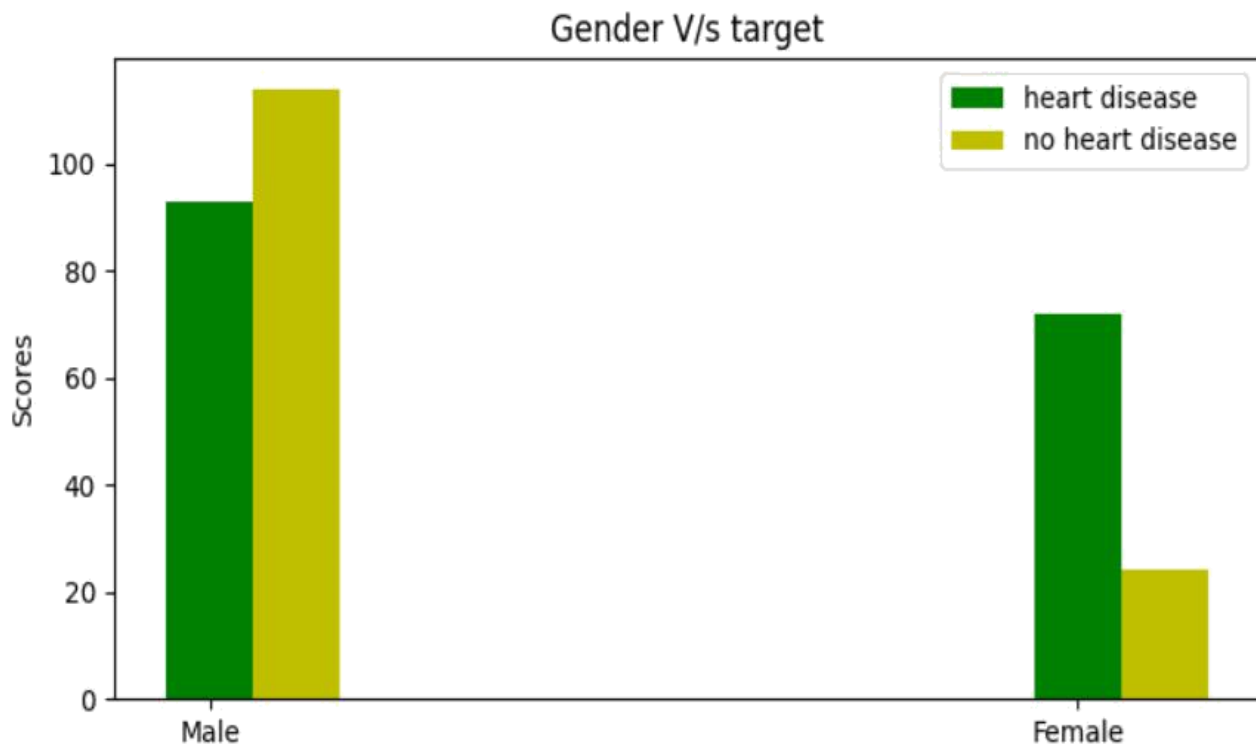


Figure 3 Balancing of Data

5.) *Disease Prediction*

Five well-known machine learning methods—SVM, Naive Bayes, Decision Trees, Random Trees, and Logistic Regression—are utilized for classification. The algorithm that has the highest accuracy is used to forecast cardiac disease after a comparative analysis of several algorithms. It is necessary to measure each algorithm's accuracy, and the algorithm with the highest accuracy is chosen to forecast heart disease. Accuracy, confusion matrix, precision, recall, and f1-score (these words are addressed in detail in the outcome section) are only a few of the assessment metrics that were considered when evaluating the experiment.

Accuracy-

Accuracy is defined as the proportion of predictions that are accurate given all the inputs to the dataset

It is written as follows:

$$\text{Accuracy} = (TP + TN) / (TP + FP + FN + TN)$$

Equation 1 Accuracy measure

6.) **Graphical Representations**

After predicting the presence of heart illness, our method offers a comprehensive graphical representation of all the data that was previously input, including both the diagnosis and the absence of a diagnostic.

The system's graphical representation demonstrates how data is input and categorized according to the existence or absence of cardiac disease. The system determines whether a person has heart disease based on entered data. Should they, the information is then added to the "Heart Disease" category. If the individual does not have heart disease, the information is added to the "No Disease" category. A comprehensive overview of the system's categorization procedure is provided by this graphical representation, which shows how data flow and classification are visually demonstrated.

5. **ML- (MACHINE LEARNING) ALGORITHMS**

A computational method known as "machine learning" includes the systematic study and analysis of algorithms that allow a system to learn from data and make forecasts without being expressly coded. It includes a variety of methods, such as unsupervised, supervised, and reinforcement learning—three separate sub-fields of the subject that let the system adapt to various types of input and feedback.

i. Logistic Regression

One of the most famous machine learning algorithms and part of the supervised learning process is logistic regression.

Use a collection of independent features to estimate the categorical dependent variable estimates the output of a categorical dependent variable from logistic regression. Therefore, results should be discrete or categorical. True or false, Yes or No, 0 or 1, etc. it could be. But instead of giving a precise value between 0 and 1, it returns a positive value between 0 and 1.

The main difference between linear regression and logistic regression is how they are used. Logistic regression is used to solve classification problems while linear regression is used to solve regression problems. In the Logistic regression, instead of a regression line, we apply a logistic function like "sigmoid" that predicts two maximums (0 or 1). The plot of the Logistic function shows that there will be many situations, such as whether the mouse is obese based on body weight or whether the cell is malignant.

Logistic regression is an essential part of machine learning given its ability to derive probabilities and classify new data using continuous and discrete data as shown in Figure 4 [11].

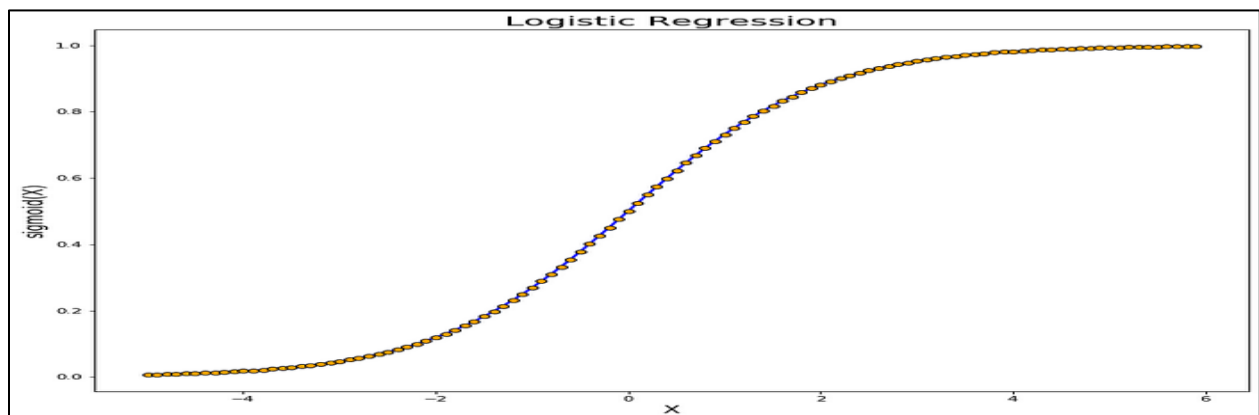


Figure 4 Logistic Regression

ii. Decision Tree

A decision tree is a machine learning algorithm that represents data in a graphical representation. The entropy of the attributes of data is used for deciding which attributes should

be used as decision nodes in the tree during the decision tree construction procedure. The entropy of a node is determined by taking the likelihood of each potential result and selecting the attribute with the greatest entropy as the tree's root. The procedure is then repeated for each branch until all nodes are allocated or the tree is complete. However, when too few or too many nodes are used, decision trees have a propensity to over-fit the data, resulting in lower precision in contrast to linear regression as shown in Figure 5. [12]

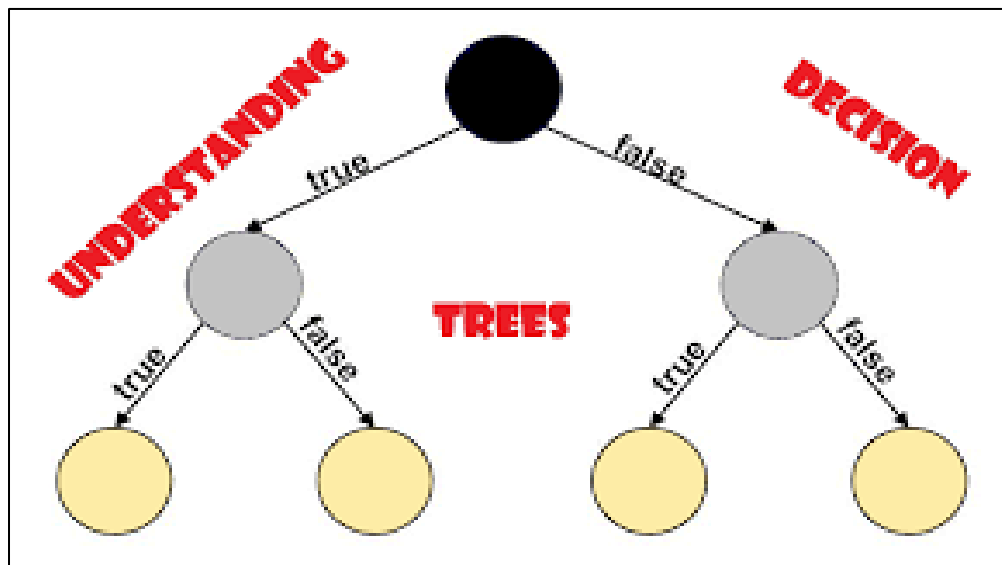


Figure 5 Decision Tree

iii. Random forest

The Random Forest classifier is a supervised machine-learning approach that may be used to solve classification and regression issues. Its foundation is the concept of ensemble learning, in which a variety of models are brought together to manage complicated issues and improve overall performance. The Random Forest approach uses a large number of decision trees that are trained on multiple data-set subsets, and the forecasts from each tree are then combined to increase the model's accuracy. Instead of relying exclusively on one decision tree, the Random Forest collects predictions from all of the trees and predicts the result based on the predictions that garnered the most votes from the participants. To improve accuracy and prevent over-fitting, the Random Forest can have more branches. The mean of all the outputs serves as the final result for classification

issues, whereas the mean of all the outputs serves as the final result for regression issues as shown in Figure 6 [13].

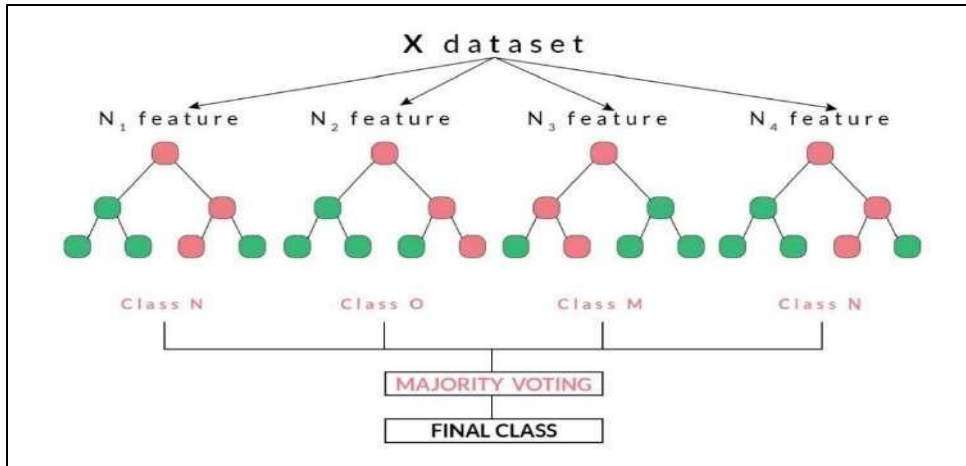


Figure 6 Random Forest

iv. Support Vector Machine

Support Vector Machine (SVM) as shown in Figure 7 [14], is a method of machine learning that uses hyperplanes to categorize data. A judgment boundary known as a hyperplane separates the groups of the data elements. The training collection consists of n vectors, designated as X_i , and matching target vectors, designated as Y_i . The kind of support vector used depends on how many hyperplanes were used for categorization. The technique is known as a linear support vector, for instance, if a line is used as the hyperplane.

#

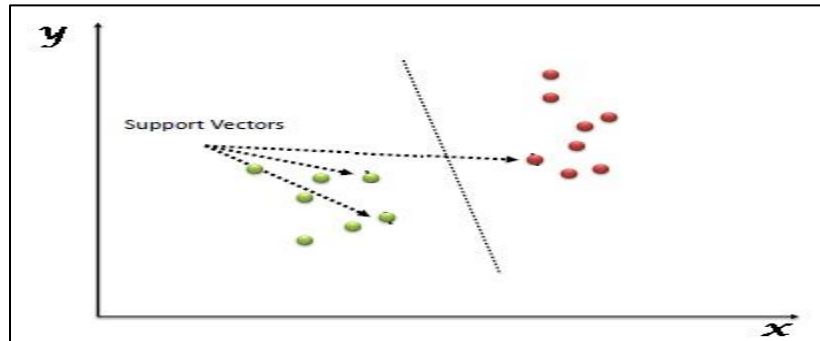


Figure 7 Support Vector Machine [14]

There are two kinds of support vector machines (SVMs):

❖ **Linear SVM:**

The Linear SVM technique is applied to data that can be split into groups by a single straight line. A dataset is said to be linearly separable in this case if it can be split into two separate groups along a straight line. The best predictor to use in these circumstances is linear SVM.

❖ **Non-linear SVM:**

Any set of data that cannot be split by a straight line is considered non-linear. Non-linear data are categorized using the Non-linear SVM algorithm. To convert the data into a higher-dimensional space, this kind of SVM uses non-linear kernel functions like polynomials, radial basis functions (RBF), or sigmoid functions. This makes it possible for the SVM to identify a hyperplane that divides the data into different groups.

v. ***K-Nearest Neighbors***

This technique, k-Nearest Neighbors (k-NN), is used to categorize data based on how far apart the data elements are from one another. One of the most important aspects of the dataset analysis is the user-defined k-fold increase in the number of adjacent data points that are taken into account for categorization. A data point is classified using the k-NN method by its k-nearest neighbors and the overwhelming class of those neighbors. There are several ways to determine the distance between two pieces of data, including the Euclidean distance, Manhattan distance, and Minkowski distance.

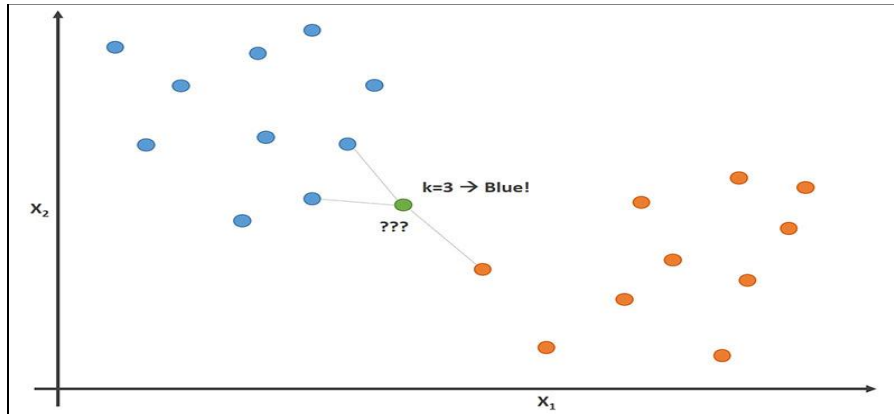


Figure 8 K-nearest neighbor

As shown in figure 8 [15], illustrates the idea of k-nearest neighbors, where $k=3$, signifying that each data point has three closest neighbors. Each cluster is indicated by a pair of two-dimensional coordinates (X_i, Y_i) , where X_i denotes the x-axis and Y_i the y-axis, and $i = 1, 2, 3, \dots, n$ denotes the number of data points in the collection.

vi. Naïve Bayes

A classification technique is the Naive Bayes algorithm used in guided machine learning. It is built on the Bayes theorem and categorizes data using a basic probabilistic method. The Naive Bayes classifier is extensively used in many areas and is regarded as one of the most simple and effective categorization techniques. It is especially helpful when dealing with big datasets because it enables the rapid creation of models for machine learning that are capable of accurate prediction as shown in Figure 9 [16].

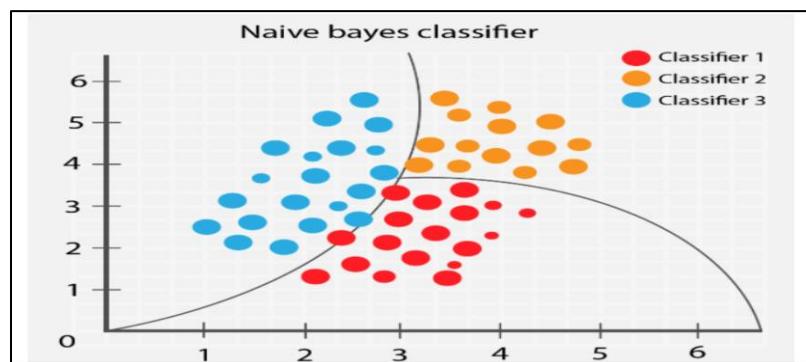


Figure 9 Naive Bayes

This technique is frequently used for large-scale data classification jobs. It uses the Naive Bayes method, a Bayes theorem-based probabilistic classifier. The classifier implies that each character in the dataset is independent of all others, which is not always the case. Despite this restriction, Naive Bayes is a fast and accurate technique for making accurate predictions with big datasets.

The following is the Bayes' theorem's formula:

$$P(A | B) = \frac{P(B | A) P(A)}{P(B)}$$

Equation 2 Bayes' theorem's formula

Where equation 2 shows:

- The posterior probability, or $P(A|B)$, measures the likelihood that a given hypothesis (A) will occur.
- $P(B|A)$ stands for Likelihood Probability, which is the likelihood that the evidence provided will result in the hypothesis being true.
- Prior Probability, or $P(A)$, is the likelihood of a theory before observation of the data.
- The probability of evidence is a marginal probability, or $P(B)$.

6. RESULT DISCUSSIONS

SVM, Naive Bayes, Decision Trees, Random Forests, and Logistic Regression are five well-known machine learning methods used in this study to predict heart disease. The heart illness UCI dataset has 76 features, however, only 14 of them are used to make predictions on heart disease. Gender, the type of chest pain, fasting blood pressure, serum cholesterol, and exang are only a few of the patient factors that were considered in this study. Each algorithm's accuracy must be evaluated, and the algorithm with the highest accuracy is selected to predict heart disease. When assessing the experiment, several assessment factors are taken into consideration, including accuracy, confusion matrix, precision, recall, and f1-score.

❖ Accuracy-

Accuracy is the proportion of predictions that are accurate given all of the inputs to the dataset.

It is expressed as:

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{FP} + \text{FN} + \text{TN})$$

Equation 3 Accuracy measure

❖ Confusion Matrix-

It provides us with a matrix as an output along with the system's overall performance.

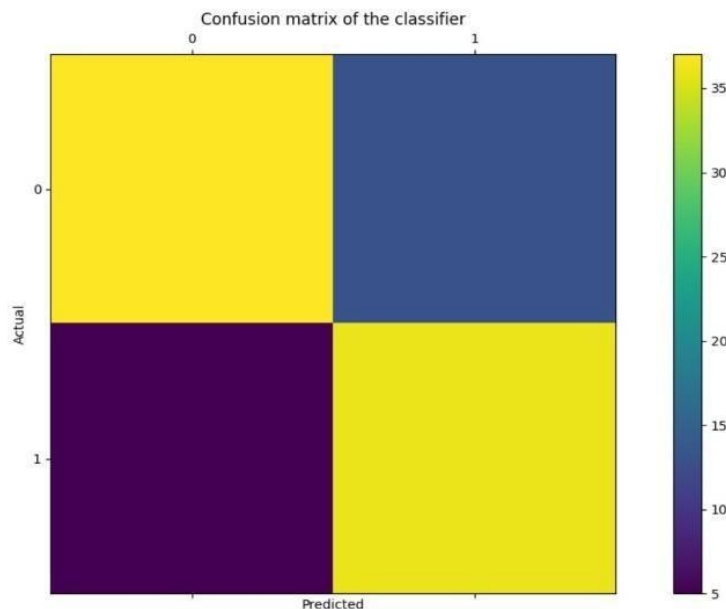


Figure 10 Confusion Matrix

Where

TP: True positive

FP: False Positive

FN: False Negative

TN: True Negative

❖ **Correlation Matrix:**

The correlation matrix is used to choose features in machine learning. It represents how various qualities are interconnected.

❖ **Precision-**

It is the proportion of correctly positive findings to all of the anticipated positive results from the algorithm.

❖ **Recall:**

Fig 10

This is the ratio of all correctly positive results to all results that the algorithm correctly predicted as being positive.

❖ **F1-Score:**

The F1 score is the harmonic mean of Precision and Recall. It rates the accuracy of the test. The range of this statistic is 0 to 1.

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{FP} + \text{FN} + \text{TN})$$

Equation 4 accuracy measurement

Here, the value has been derived from the number count of TPs (true positive), TNs (true negatives), FPs (false positives), and FNs (false negatives) using Equation 4. SVM is shown to be the best with 85% accuracy.

In Figure 12, GUI is created using the Tkinter Python package. The figure displays the 14 attributes that patients must enter into the appropriate boxes. Following the summary of



Figure.11 Correlation Matrix

the information, the patient's entered details are displayed, and the patient's results are displayed.

In the sections that follow, we will summarize accuracy. Figure 10 and Table 2 further present accuracy comparisons of many well-known machine learning models that we used in our study effort.

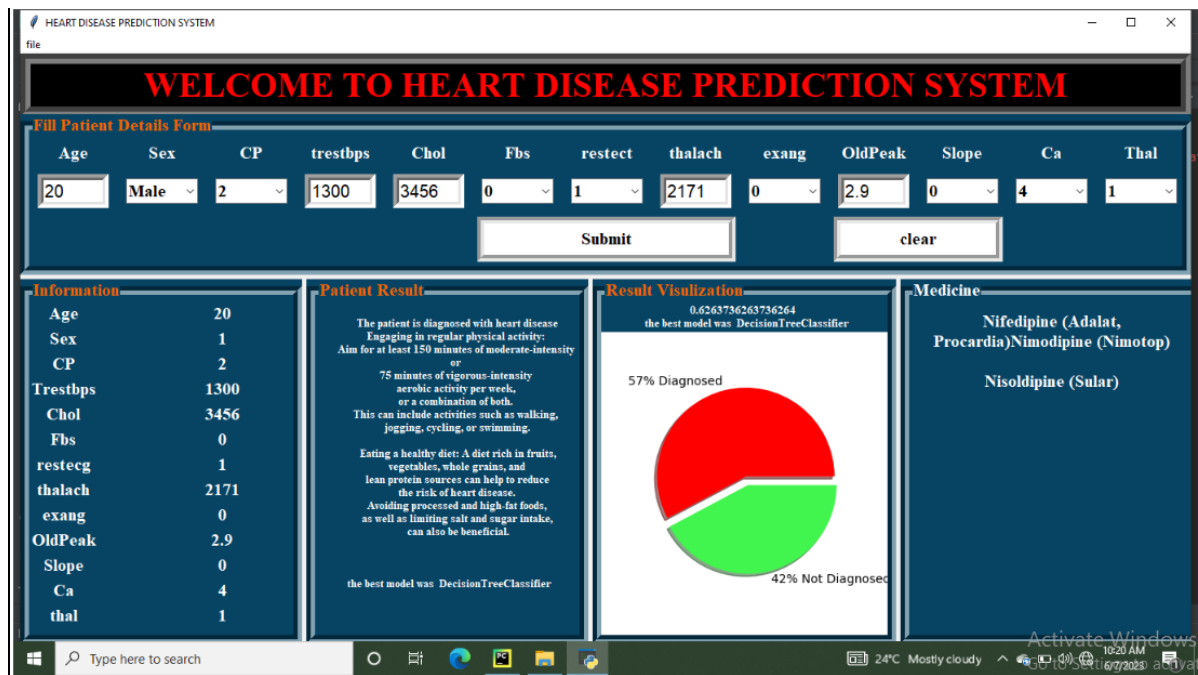
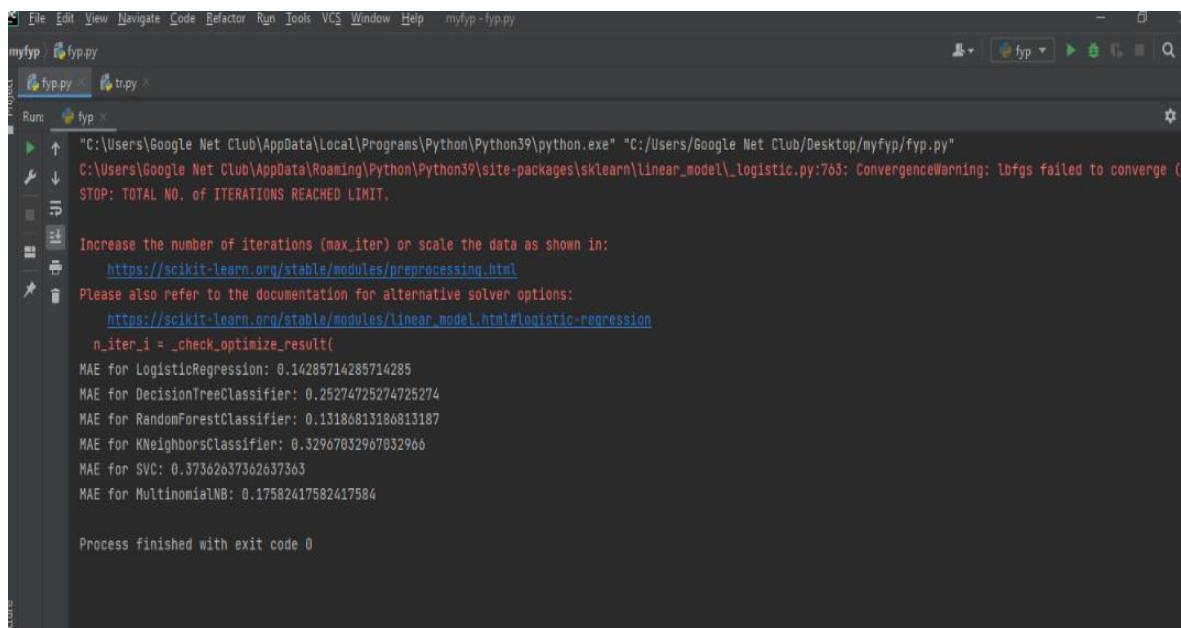


Figure 12 Input & output scree

Comparison of algorithms' accuracy Algorithm Accuracy

The accuracy comparison of the five well-known algorithms of machine learning is shown below, the accuracy is measured using the above equation 2 of all the algorithms as shown in the below figure.



```

"C:\Users\Google Net Club\AppData\Local\Programs\Python\Python39\python.exe" "C:/Users/Google Net Club/Desktop/myfyp/fyp.py"
C:\Users\Google Net Club\AppData\Roaming\Python\Python39\site-packages\sklearn\linear_model\_logistic.py:763: ConvergenceWarning: lbfgs failed to converge (st
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
https://scikit-learn.org/stable/modules/linear\_model.html#logistic-regression
n_iter_i = _check_optimize_result(
MAE for LogisticRegression: 0.14285714285714285
MAE for DecisionTreeClassifier: 0.25274725274725274
MAE for RandomForestClassifier: 0.13186813186813187
MAE for KNeighborsClassifier: 0.32967032967032966
MAE for SVC: 0.37362637362637363
MAE for MultinomialNB: 0.17582417582417584

Process finished with exit code 0

```

Figure 13 accuracy

Table 2 Accuracy Table

Algorithm	Accuracy
Support Vector Machine (SVM)	81.2%
Logistic Regression(LR)	79.1%
Random Forest(RF)	79.1%
Naive Bayes(NB)	76.9%
Decision Tree(DT)	85%

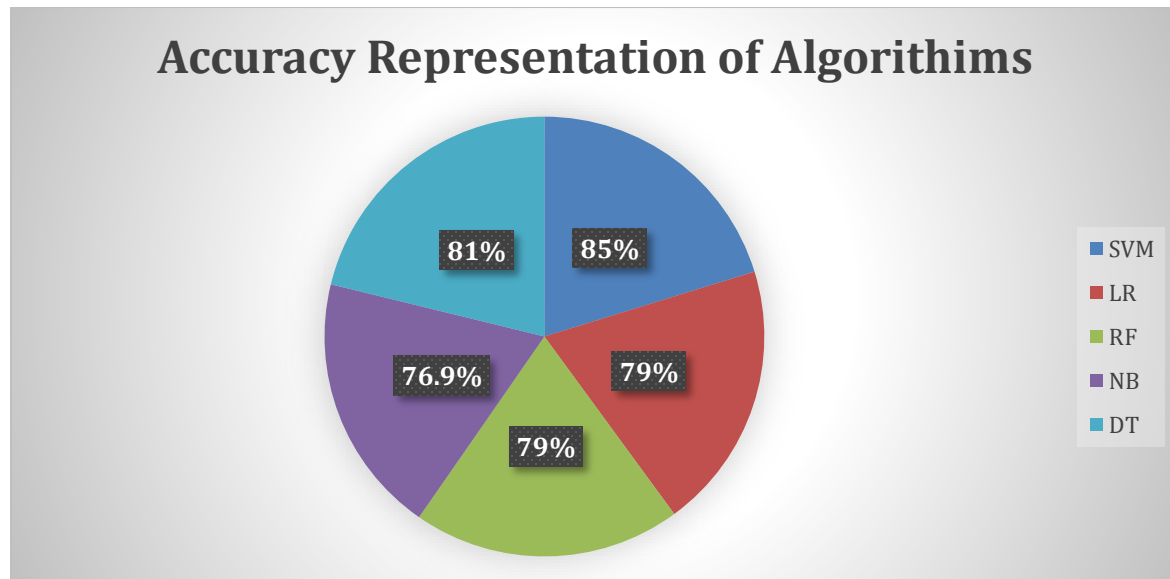


Figure 14 Accuracy Representation of Algorithms

7. CONCLUSION & FUTURE WORKS

A huge social impact might result from the earliest prediction of cardiac problems using cutting-edge technology like machine learning. Not just in Pakistan but also globally, heart disorders continue to be the biggest cause of death. This technology technique can enable early diagnosis of heart problems, empowering high-risk patients to decide on lifestyle changes that can reduce consequences. This is a major development in the medical profession. Early detection and care are urgently needed since cardiac disorders are being discovered in an increasing number of patients each year. Both the medical industry and patients may greatly benefit from the incorporation of appropriate technology help in this area.

The five most famous machine learning algorithms, including Support Vector Machines (SVM), Decision Trees, Random Forests, Naive Bayes, and Logistic Regression, were evaluated using the data-set in this study.

The data-set includes 76 distinct features linked to variables that affect patients' risk for heart disease. 14 important variables are chosen from among these traits as being essential for evaluating the system. When taking into account all 76 aspects at once, the author notices a loss in system efficiency. A thorough selection of qualities is made to improve efficiency. The goal is to determine the ideal number of attributes (n) that produce a more precise model. Some data-set properties are disregarded during attribute selection because of their strong association with other

variables. Efficiency is improved significantly by omitting the inclusion of all data-set characteristics.

A prediction model was created after the five distinct most famous machine learning methods' effectiveness was assessed. The main objective of this model is to precisely forecast the presence of a certain illness. Several assessment tools, including the confusion matrix, accuracy, precision, recall, and f1-score, are used to do this. The decision tree classifier outperforms the other five classifiers in terms of assessment metrics, with an accuracy of 85%.

References

1. "<https://mmi.edu.pk/blog/heart-attack-cases-in-pakistan/>".
2. "link," <https://www.analyticsvidhya.com/blog/2021/03/everything-you-need-to-know-about-machine-learning/>.
3. S. a. S. R. Chellammal, "Recommendation of attributes for heart disease prediction using correlation measure," *Int J Recent Technol Eng (IJRTE)*, vol. 8, pp. 870--875, 2019.
4. H. a. A. S. a. K. R. a. J. R. a. N. P. Jindal, "Heart disease prediction using machine learning algorithms," *IOP conference series: materials science and engineering*, vol. 1022, p. 012072, 2021.
5. S. a. T. C. a. S. G. Mohan, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542--81554, 2019.
6. U. a. S. D. a. C. P. Nagavelli, "Machine learning technology-based heart disease detection models," *Journal of Healthcare Engineering*, vol. 2022, 2022.
7. A. A. M. S. D. R. a. D. P. G. vii. Apurb Rajdhan, "Heart Disease Prediction using Machine Learning," *International Journal of Engineering Research & Technology (IJERT)*, vol. 9, no. 04, April 2020.
8. J. P. a. H. A. U. a. D. S. U. a. K. J. a. K. A. a. S. A. Li, "Heart disease identification method using machine learning classification in e-healthcare," *IEEE Access*, vol. 8, pp. 107562--107582, 2020.
9. V. a. D. A. a. R. M. K. Ramalingam, "Heart disease prediction using machine learning techniques: a survey," *International Journal of Engineering & Technology*, pp. 684--687, 2018.
10. "uci," <https://archive.ics.uci.edu/ml/datasets/Heart+Disease>.

11. "Linear regression," <https://towardsdatascience.com/logistic-regression-explained-9ee73cede081>.
12. "DT," <https://www.linkedin.com/pulse/intuition-decision-tree-ensemble-regression-ananalysis-gunji>.
13. "RF," [Online]. Available: https://www.google.com/imgres?imgurl=https%3A%2F%2Fupload.wikimedia.org%2Fwikipedia%2Fcommons%2F7%2F76%2FRandom_forest_diagram_complete.png&tbnid=6sr7Swk5a-xyVM&vet=12ahUKEwjS4YyhtsT_AhXQpicCHYPIC3gQMygBegUIARDhAQ..i&imgrefurl=https%3A%2F%2Fen.wikipedia.o.
14. "SVM," [Online]. Available: <https://www.spiceworks.com/tech/big-data/articles/what-is-support-vector-machine/>.
15. "KNN," [Online]. Available: <https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning>.
16. "naive Bayes," [Online]. Available: <https://www.analyticsvidhya.com/blog/2022/03/building-naive-bayes-classifier-from-scratch-to-perform-sentiment-analysis/>.
17. "decision tree," <https://medium.com/analytics-vidhya/understanding-decision-tree-3591922690a6>.
18. K. a. S. R. a. o. Saxena, "Efficient heart disease prediction system," *Procedia Computer Science*, vol. 85, pp. 962--969, 2016.
19. S. A. K. A. Nikhar, "Prediction of heart disease using machine learning algorithms," *International Journal of Advanced Engineering, Management and Science*, vol. 2, p. 239484, 2016.
20. <https://my.clevelandclinic.org/health/diseases/21493-cardiovascular-disease>