Heart Disease Prediction Using MachineLearning

PROF. G.R. RAO[1], SHOURYA KHUJNERI[2], ATUL KUMAR TOMAR[3], RUDRAKSH SHARMA[4]

[1] *Associate Professor, Department of computer Engineering Bharati Vidyapeeth Deemed to beUniversity College of Engineering, Pune, India*

[2, 3, 4] *Department of computer Engineering Bharati Vidyapeeth Deemed to be University College of Engineering, Pune, India*

Abstract— The contents of this study primarily focus on various data mining approaches that are useful in predicting heart disease using various data mining technologies that are available. The brain, kidneys, and other areas of the human body will suffer if the heart does not operate properly. Heart disease is a condition that impairs the heart's ability to operate. In today's world, heartdisease is the leading cause of mortality. This study gathered data from a variety of sources and dividedit into two parts: 80 percent for the training dataset and 20% for the test dataset. Different classifier methods were used to improve accuracy, which was then summarised. Random Forest Classifier, Decision Tree Classifier, Support Vector Machine, k-nearest neighbour, Logistic Regression, and Naive Bayes are the methods in question. SVM, Logistic Regression, and KNN all performed aswell as or better than other methods. This research offers a development in which fundamental prefixes such as sex, glucose, blood pressure, heart rate, and others are used to determine whichfactors are prone to heart disease. The paper's next aim is to conduct real-life tests using various equipment and clinical trials.

Indexed Terms-- Machine Learning, Decision Tree Classifier, Random Forest Classifier, SVM, Logistic Regression, k-nearest neighbor

I. INTRODUCTION

Heart disease is now a widespread and prevalent illness in the human body, and it has claimed the lives of many people throughout the world. According to the World Health Organization, 12 million people die each year from heart disease [2]. Cardiovascular, heart attack, coronary, and knock are examples of heart illnesses. Knock is a kind of cardiac disease that is caused by the strengthening, blockage, or narrowing of blood arteries that go through the brain, or by high blood pressure. The superiority of facilities is the biggest difficulty that the healthcare business faces today. The quality of service willbe determined by accurately diagnosing the ailment and delivering appropriate therapy to patients.

A poor diagnosis can has severe implications. Medical history records or data are extensive, yet they come from a variety of sources. Physician interpretations are critical components of these data. Because real-world data is likely to be noisy, partial,and inconsistent, data pre-processing will be necessary in the directive to fill the database's missing values.

Even though cardiovascular illnesses were formerly a major cause of mortality in the globe, they have now been declared themost preventable and controllable diseases. The entire andproper care of a disease is dependent on the disease's timely judgement. A proper and thorough approach for identifying high-risk individuals and mining data for rapid heart infection analysis appears to be a critical need. Distinct people's bodies exhibit different heart disease symptoms, which might vary.

II. TECHNOLOGY CLASSIFIERS

- PANDAS: - Pandas is a data manipulation and analysis software package for the Python programming language. It provides data structures and functions for manipulating numerical tables and time series in particular. It's free software distributed under the BSD three- clause licence.
- SEABORN: - Seaborn is a Python data visualization package that uses matplotlib as its foundation. It also has a high-level interface for creating visually appealing and useful statistics

visuals.

- PLOTLY: - The plotly Python library is an interactive, opensource plotting framework that supports over 40 different chart types for statistical, financial, geographic, scientific, and 3-dimensional applications.

- SKLEARN: - Scikit-learn (previously scikits. Learn and also known as sklearn) is a Python machine learning package. It includes support-vector machines, random forests, gradient boosting, k-means, and DBSCAN, among other classification, regression, and clustering techniques, and is designed to work with Python numerical and scientific libraries NumPy and SciPy. Scikit-learn is a NumFOCUS-supported project.

III. PROPOSED METHODOLOGY

A. Data Collection – heart disease data is collected fromKaggle [3]

S. No	Feature	Abbreviation
1.	Age	Age in years
2.	Sex	(1 = male; 0 = female)
3.	Cp	chest pain type
4.	trestbps	resting blood pressure
5.	chol	serum cholesterol in mg/dl
6.	fbs	(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false) '>126' mg/dL signals diabetes
7.	Restecg	resting electrocardiographic outcomes 0: Nothing to note
8.	thalach	The maximal heart rate was reached.
9.	Exang	ST depression caused by activity compared to rest examines cardiac stress.
10.	Oldpeak	ST depression induced by exercise relative to rest looks at stress of heart
11.	Slope	the slope of the peak exercise ST segment
12.	Ca	number of major vessels (0-3) coloured by flourosopy
13.	Thal	thallium stress result
14.	Target	have disease or not

Table of features

Each column has a distinct feature that affects heart diseaseeither directly or indirectly. So I utilised all of the qualities at first, and then, depending on risk factors and prediction, I lowered the attribute to improve performance and eliminate a key risk factor This data collection includes. The remainder of the feature is merely one target value. Age and gender are allocated to features, with 0 and 1 for gender respectively, maleand female

B. Data Pre-processing –The dataset had a lot of NULL and missing values. These were eliminated using a variety of methods, including replacing missing data with the mean, median, and mode of the existingvalues. A correlation matrix is used to determine the relationship between the existing characteristics. Positively corelated numbers are closer to 1, while negatively correlated values are closer to –1.

C. Algorithms and Model Description –For prediction tasks, several Supervised Machine Learning models are utilised. The following are the various machine learning algorithms that were employed for the task:-

- Logistic Regression: A key classification approach is logistic regression. It is comparable to polynomial andlinear regression and belongs to the category of linearclassifiers [5]. Logistic regression is a simple and quickmethod of analysing data, and the findings are easy tounderstand. Although it is primarily a binary classification approach, it may also be used to solve multiclass issues.

- Implementation

The implementation steps for logistic regression are given a follow:

a) *Obtain the probabilities Mapping predicted values toprobabilities, using the Sigmoid function.*

$1(1 + e^{-y})$

Where, y is input to the function and e is the base of natural log

Obtain the probabilities by following equations:

$P = e^{y} / 1 + e^{y}$

here P is the probability of success. The eqn is the Logic Function

q is the probability of failure written as:

$q = 1 – P= 1 – (e^{y} / 1 + e^{y})$ (8) where q is the probability of failure On dividing we get

$$\frac{p}{1-p} = ey$$

On taking log on both sides,

Log ^{p} =y

$1-p$

Here (p/1-p) is the odd ratio. When the „y" is positive, the probability success is more than 50%.

b) Decision Boundary- Probability to Class Mapping is a decision boundary that maps probabilities to classes. The probability score returned by the prediction function is between 0 and 1. To classify something intoa discrete category, a threshold value is chosen abovewhich it is categorised as class 1 or class 2. If our threshold was 0.5 and our function value was 7, it would be considered positive. For example, if the value is.3, the categorization is negative. Multiple classes can be used in logistic regression, with the highest probability predicted class being taken into account.

- K-nearest neighbors (KNN): The KNN method is a supervised machine learning technique that may be used to handle classification and regression issues[6].It's simple to set up and comprehend, but it has the disadvantage of being substantially slower as the amount of data in use rises.

- Support Vector Machine: SVM stands for Support Vector Machine and is one of the most widely used Supervised Learning algorithms for Classification and Regression issues. However, it is mostly utilized in Machine Learning for Classification difficultiess. The SVM algorithm's purpose is to find the optimum line or decision boundary for categorizing n-dimensional space into classes so that additional data points may bereadily placed in the proper category in the future. A hyperplane is the name for the optimal choice boundary.

- Implementation

The following is a description of how Support Vector Machine is implemented:

a) Load the data sets and clean values, in case of no value for a particular feature in a row replace with themedian value the row from the dataset.
b) S In a 60:40 ratio, divide the data set into train and test groups.
c) Selecting a Linear Kernel Function or a Radial Basis Function as the Kernel Function.
d) Creating a hyper plane with the assistance of a test data set before using SVM.

i. T The train data is taken and both Kernel function namely Linear Kernel Function or Radial Basis Function is applied.
ii. A Apply test data set on the trained model.
iii. T The model uses hyper plane and finds closest proximity to either class that is having heart disease(yes/1) or not having heart disease (no/0).

e) Calculate the accuracy using:

$$\text{Accuracy} = \frac{\text{(number of data items pridicted = actual value in test data set}}{\text{total number of values in test data set}}$$

- Random forests: Random forests, also known as random choice forests, is an ensemble learning approach for classification, regression, and other problems that works by training a large number of decision trees. For classification tasks, the random forest's output is the class chosen by the majority of trees.

- Implementation

The following is a description of how random forest is implemented:

a) Load the heart disease dataset.
b) Using the Random Forest Classifier tool, split the heart disease dataset into train and test data with a 60:40 percentage.
c) K-Fold When a given knowledge set is divided into a K range of sections/folds, each fold is used as a testing set for some purpose, this is known as cross validation.
d) Use a train set to train the model.
e) Predict how the test fold will turn out.
f) Connect the dots between expectations and outcomes (only possible outcomes are 1 and0)
g) Calculate the accuracy.

$$\text{Accuracy} = \frac{\text{(TP+TN)}}{\text{(TP+TN+FN+FP)}} \times 100$$

Where,

TP- True Positive (prediction is yes, andthey do have

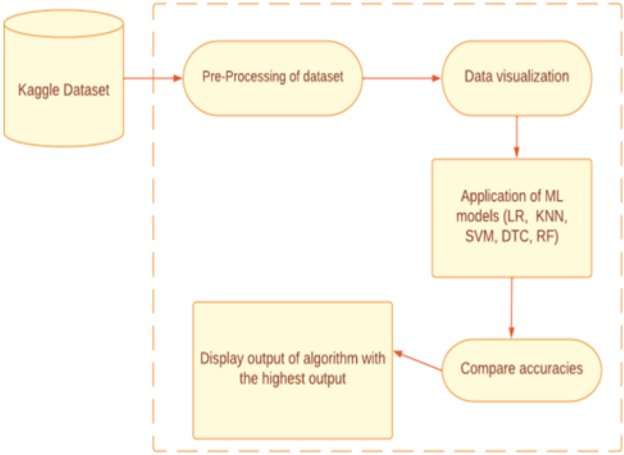
the disease.
TN-True Negative (prediction is no, and they don't have the disease.)

FP-False Positive (We predicted yes, butthey don't actually have the disease. (Also known as a "Type I error.") FN-False Negative (We predicted no, but they actually dohave the disease. (Also known as a "Type II error.")

The accuracy obtained by using random forest algorithm is81.31%.

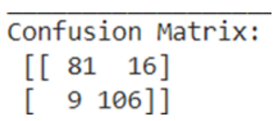
- Logistic regression hyperparameter tuning: - Hyperparameters in machine learning algorithms allow you to adjust the algorithm's behavior to your individual dataset. Hyperparameters are not the same as parameters, which are the internal coefficients or weights discovered by the learning procedure for a model.

D. System Architecture


Kaggle Dataset
Pre-Processing of dataset
Data visualization
Application of ML models (LR, KNN, SVM, DTC, RF)
Compare accuracies
Display output of algorithm with the highest output

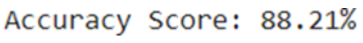
IV. RESULT AND DISCUSSION

Confusion Matrix: - Important predictive metrics like recall, specificity, accuracy, and precision are shown using confusion matrices. Confusion matrices are valuable because they allow you to compare values such as True Positives, False Positives, True Negatives, and False Negatives directly. The confusion matrix is a matrix that is used to evaluate the classification models' performance for a given set of test data. Only if the realvalues for test data are known can it be determined.

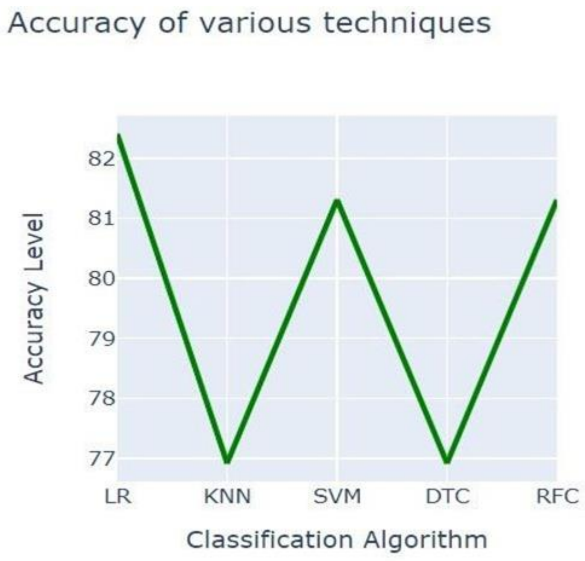


```
Confusion Matrix:
 [[ 81  16]
 [  9 106]]
```

Accuracy: Accuracy = (True Positive+ True Negative)/ Total



```
Accuracy Score: 88.21%
```



- Precision: Precision is defined as the percentage of accurately anticipated positive instances divided by the total number of expected positive cases [4]. Belowthis bar chart indicates the precision of all algorithms.

Precision = True Positive/ (True Positive + False Positive)

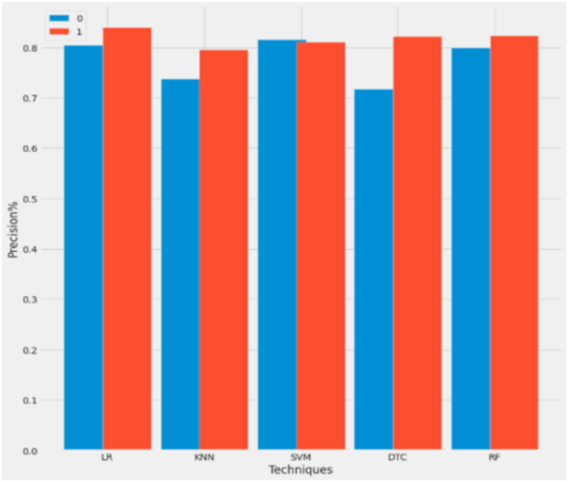


Fig: Precision of various techniques.

- Recall: Recall is the proportion of accurately predicted positive cases in the base class to all cases in the base class.

Recall = True Positive/ (True Positive + False Negative)

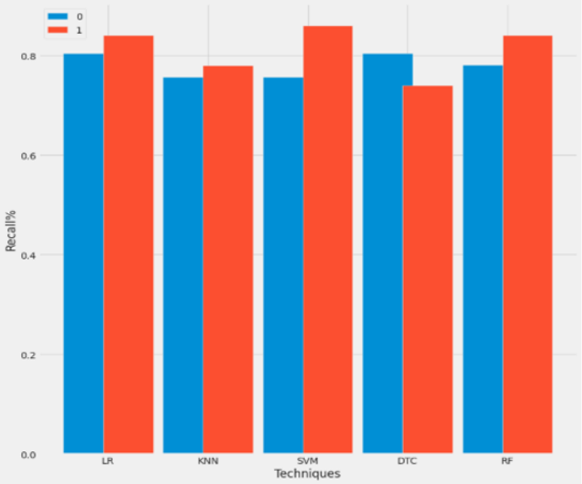


Fig: Recall of various techniques.

- F1 Measure: -The F1 measurement is the average accuracy and recall. As a result, this test included bothfalse negative and false positive results. F1 is more helpful than precision. Accuracy is better when both false positive and false negative costs are comparable,yet F1 is good for quite different scenarios.

F1 score = (2 * Precision * Recall))/ (Precision + Recall)

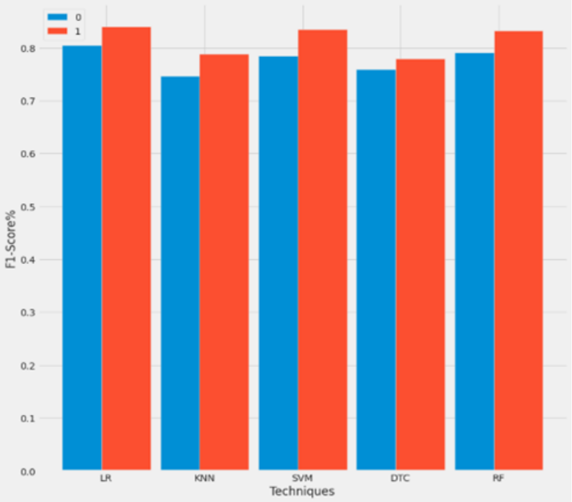


Fig: Fl-Score of various techniques.

S.NO	Models	Accuracy	Precision	Recall	F1 score	Support
1	Logistic Regression	82.42	0.82	0.82	0.82	91.00
2	K-Nearest	76.92	0.76	0.76	0.76	91.00
3	Support Vector Machine	81.32	0.81	0.81	0.81	91.00
4	Decision Tree Classifier	76.92	0.77	0.76	0.77	91.00
5	Random Forest	81.32	0.81	0.81	0.81	91.00

Table: Results of various ML models.

V. FUTURE SCOPE

Many smart devices and their applications in the healthcareindustry are all too familiar to us these days. These devices maybe used to calculate walking time, workout pace, and other activities. As a result, smart devices may be used to predict cardiac illness and offer data for these digital advantages. We may also use heart disease methods to predict cancer, diabetes,and other diseases, and we can use a new algorithm to increaseaccuracy and performance.

VI. CONCLUSION

Finally, we'll say that this project Disease prediction is extremely useful in everyone's day-to-day lives, but it's especially important for the healthcare sector because they're the ones who use these systems daily to predict patients' diseases and support their general information and symptoms. Nowadays, the health industry plays an important role in curing patients' diseases, so it's often quite useful for the health industry to inform the user, and it's also quite useful for the userjust in case he or she doesn't want to travel to the hospital or other clinics, so simply by entering the symptoms and any otheruseful information within the form, the user can get to know thedisease he or she is affected by, and the health industry can benefit as well. If the health sector embraces this idea, doctors' workload will be decreased, and they will be able to forecast the patient's sickness more readily. Disease prediction is the provision of predictions for a variety of common diseases that,if left untreated and often neglected, can progress to severe diseases and create a slew of issues for the patient. This project may be updated in the future by adding additional attributes to the dataset and making it more interactive for the users. It can also be done as a mobile application. We'll make changes to the system by linking it to the hospital's database.

The study finishes with the use of multiple Machine learning classification algorithms to determine whether a person has heart disease. The classification methods Logistic Regression, Support Vector Machine, and Random Forest Classifier perform better than the other two, with K-Nearest Neighbours Algorithm providing the poorest accuracy.

Furthermore, among the five ML Models employed, Logistic Regression performs the best and may be utilised to design a system that can assist patients determine whether they have heart disease.

ACKNOWLEDGMENT

The project and paperwork could not have been possible without our guide, Professor G.R Rao, who helped us by providing the necessary guidance and resources required for the paperwork.

REFERENCES

[1] L. Ali et al., "An Optimized Stacked Support VectorMachines Based Expert System for the Effective Prediction of Heart Failure," IEEE Access, vol. 7, pp. 54007-54014, 2019, doi: 10.1109/ACCESS. 2019.2909969.

[2] A. Javeed, S. Zhou, L. Yongjian, I. Qasim, A. Noor, and R. Nour, "An Intelligent Learning System Based on RandomSearch Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection," IEEE Access, vol.7, pp. 180235-180243, 2019, doi: 10.1109/ACCESS.2019.2952107.

[3] Fares sayaha: "//www.kaggle.com/code/faressayah/predicting-heart- disease-using-machine-learning/notebook"

[4] Koo Ping Shung," Error! Hyperlink reference not valid."

[5] R. Detrano et al., "International application of a new probability algorithm for the diagnosis of coronary arterydisease," Am. J. Cardiol., vol. 64, no. 5, pp. 304-310, 1989.

[6] Beant Kaurh, Williamjeet Singh, "Review on Heart Disease Prediction System using Data Mining Techniques", © IJRITCC, Vol.2, Issue: 10, p.p.3003-08,2014.

[7] Rezaul Karim, Mohaiminul Islam, Wang Chengliang "Comparative Study to Identify the Heart Disease Using Machine Learning Algorithms"