*International Journal of Research in Engineering and Science (IJRES)*
*ISSN (Online): 2320-9364, ISSN (Print): 2320-9356*
*www.ijres.org Volume 11 Issue 3 ǁ March 2023 ǁ PP. 592-601*

# Prediction of Cardiovascular Disease UsingMachine Learning and Data Science.

Pragya Tiwari
*Department of CSE*
*Galgotias University, greater Noida(India)*

Shivang Singh
Department of CSE
*Galgotias University, greater Noida(India)*

Jatin Singh
*Department of CSE*
*Galgotias University, greater Noida (India)*

Divyansh
*Department of CSE*
*Galgotias University, greater Noida(India)*

***Abstract***
*Conditions affecting the heart are often referred to as "heart disease." According to the latest data from the World Health Organization, cardiovascular illnesses remain the leading cause of mortality globally, accounting for 17.9 million fatalities per year.*
*Heart Disease is the most dangerous life-threatening chronic diseaseglobally.*
*It's common knowledge these days that artificial intelligence and machine learning are having a profound impact on the healthcare sector. Predicting Cardiovascular Diseases using Machine Learning Algorithms. Obtaining a patient's medical history is easybecause to the abundance of freely available online resources.*
*The work's goal is to use Machine Learning techniques like KNN, Random Forest, Logistic Regression, etc., to forecast whether or not a certain patient would develop heart disease.*
*We will first collect, prepare and clean the dataset which is the first step for any data science project. We would using Python libraries like Numpy and Panda for Data preparation and cleaning.*
*We will also visualize our dataset through various data visualization plots . Data visualization would be achieved with the help of Pyplot and Sklearn.*
*After training and successfully running our ML models we would calculate the efficiencies of all the ML algorithms used. Finally we will compare all the models using confusion matrix and classificationreport to get to know which ML algorithm is the best. Given a dataset, we will predict whether a given person is suffering from heart disease or not.The dataset will contain information like age, sex, chest pain, blood pressure, cholesterol level etc.*
*Consequently, we have included cardiovascular disease and its lifestyle factors as well as machine learning methodologies in this research. We have predicted cardiovascular disease using these machine learning methods and analysed and compared the various machine learning algorithms that were utilised in the experiment to make this prediction. The purpose of this study is to investigate the feasibility of using machine learning to predict cardiac events.*

---

Date of Submission: 15-03-2023 Date of acceptance: 30-03-2023

---

## I. Introduction

[1] A heart attack occurs when the heart's blood flow is severely reduced or stopped altogether. A buildup of fatty deposits, cholesterol, and other substances in the heart's (coronary) arterial occlusion is caused by. Deposits of cholesterol and fatty substances are called plaques. The buildup of plaque in artery walls is medically known as atherosclerosis.

Sometimes a plaque may rupture, causing a clot to form and so reducing blood flow. Reduced blood supply to the heart may cause damage or even death to the heart muscle. A heart attack is also known as a

myocardial infarction. Heart attack symptoms can vary. Mild symptoms are present in some people. Others display serious symptoms. Some individuals show no symptom.

[2] Common Typical signs of a heart attack include:

- Chest pain, which may feel like stress, stiffness, physical discomfort, squeezing, or hurting; discomfort or pain that extends to the shoulder, forearm, rear, throat, jaws, teeth, andeven the upper abdomen; Fatigue.
- Symptoms of acid reflux or indigestion
- Problems with dizziness or loss of consciousness that comeon suddenly Nausea

Pain that develops rapidly in the back, arms, or neck is an example of a symptom that is more unusual in females. Somepeople experience sudden cardiac arrest as the first sign of a heart attack.

Sudden cardiac arrests are possible. But many individuals have symptoms and signs seconds, weeks, or even weeks beforehand. [3]Signs of angina include discomfort in the chest that lasts longer than a few minutes of rest and becomes more difficult to ignore. Angina is brought on by a brief decrease in the heart's blood flow. [4] Since the advent of the digital era, massive volumes of information have been gathered and stored. A plethora of information is being acquired since monitoring as well as other data collecting technologies are readily accessible and routinely employed in today's hospitals. Machine learning is commonly employed nowadays to evaluate this data and detect issues in the healthcare profession since it is very difficult, if not impossible, for humans to do so. [7] When applied to big data sets, reinforcement learning (ML) has shown to be an excellent decision-making and prediction tool.

quantity amount of information the healthcare sector generates. We have also witnessed the use of ML methods inrecent innovations across several IoT domains (IoT).

Predicting cardiac disease using ML approaches is onlysomewhat explored in the existing literature.

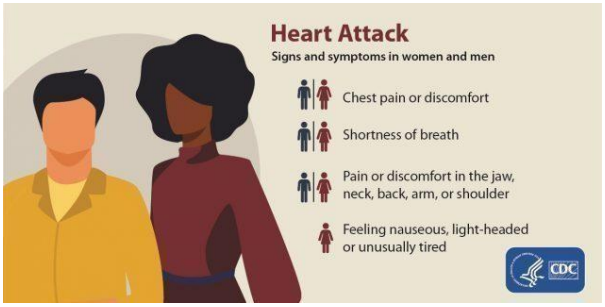



Figure 1.

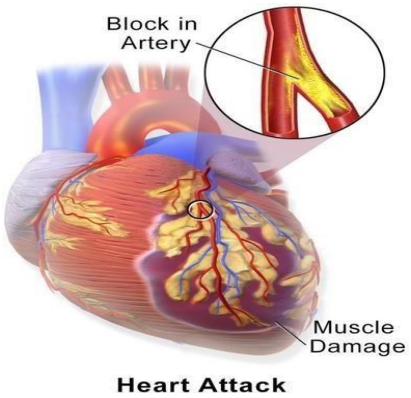



Figure 2.

[2] The heart is the most important organ because it pumps blood throughout the body, the human heart is considered to be the most vital organ. In addition to the rib cage, the heart is shielded by double-layered tissue membranes.

The heart is a four-chambered structure that filters oxygenated and unoxygenated blood. The human heart is roughly the size of a fist and contains all five kinds of blood vessels (arteries, veins, capillaries, arterioles, and venules).

**DATA DESCRIPTION**

[3] Heart disease is caused not just by hereditary risk factors like high blood pressure and diabetes, but also by environmental ones like Thalach, Exang, Oldpeak, Slope, Ca, Thal, Nun. In addition to these, other key risk factors includedietary choices, lack of physical exercise, and excess body fat.

Heart disease is exacerbated by diabetes for a number of other reasons. Heart disease is caused by a number of factors, including smoking, which increases the risk of developing heart problem, high vital sign, which causes the heart work harder to pump the blood and may strain the hearts and damage capillaries, and abnormal cholesterol levels.

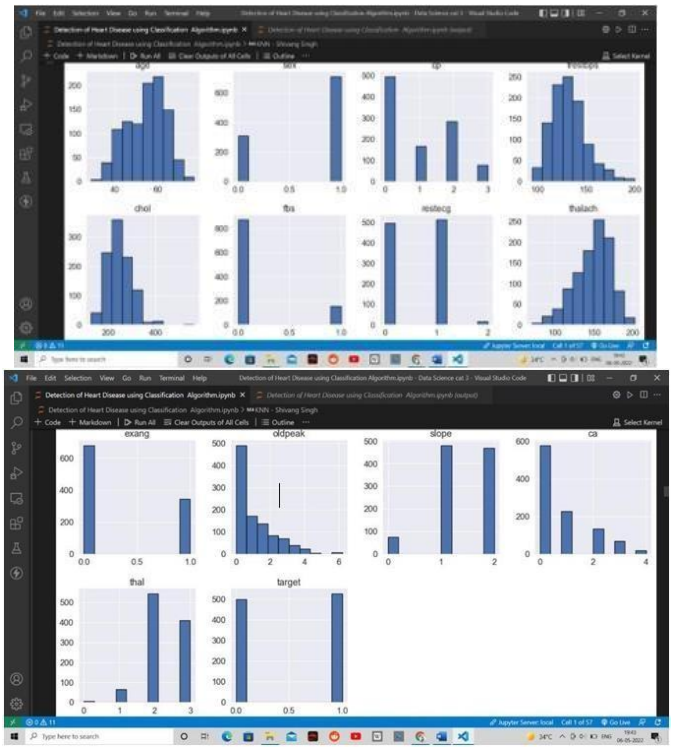
On one data set, we ran a computer simulation. Our dataset is a csv dataset named heart.csv. The dataset contains 1025 samples (rows) and 14 input features (columns). The features describe Health records of various patients. The output featuretells whether a person has a heart disease or not.

A list of all features is given in Table .

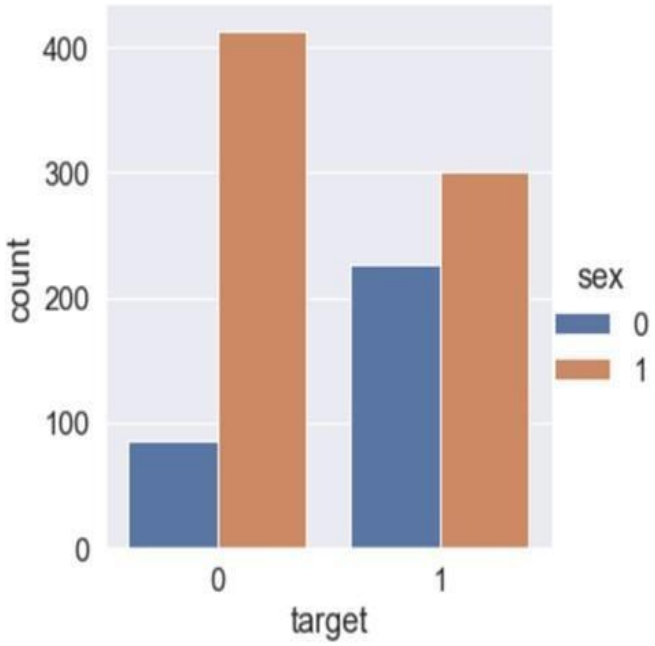
| | |
|---|---|
| age | In years, this is the patient's age. |
| sex | Individual's gendered status as a patient. A man is indicated by a 1, and a female by a 0. |
| cp | — Asymptomatic (Type 0) is the chest pain category with the lowest value.<br>— The first value describes atypical angina,<br>— whereas the second describes discomfort that is not related to the heart.<br>— Third value: regular angina |
| trestbps | BP at rest (mm Hg on admission to the hospital.) |
| chol | Total cholesterol levels as measured in milligrammes per deciliter. |
| fbs | Individual's fasting blood sugar level (> 120 mg/dl, 1 = true; 0 = false). |
| restecg | — Using the criteria established by Estes, a resting electrocardiogram with a value of 0 indicates the presence of left ventricular hypertrophy, whereas a value of 1 indicates normality.<br>— ST-T wave affliction (T wave rearrangements and/or ST spike or depress of > 0.05mV) is a value 2 condition. |
| thalach | The fastest a person's heart can beat. |
| exang | Angina during exercise (1 for affirmative, 0 for no) |
| oldpeak | Reduced ST segment ('ST' refers to locations on the ECG plot) as compared to resting ST segment. |
| slope | The slope of the peak exercise ST segment — 0: downsloping; 1: flat; 2: upsloping It's downsloping if the value is 0,flat if it's 1, and upsloping if it's 2. |
| ca | How many large ships there there (0–3). |
| thal | Thalassemia, a genetic blood condition Value 0: not used (already removed from dataset). One: a malfunction that has been repaired (no blood flow in some part of the heart)<br>Standard blood flow (2nd value)<br>(3) Reversible flaw (a blood flow is observed but it is not normal) |
| target | Coronary artery disease (may be 0 or 1) |

Table 1. Detail of Dataset

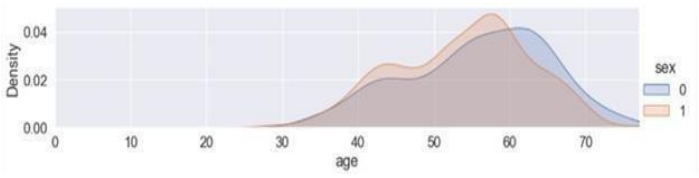
**Visualizing our Dataset**

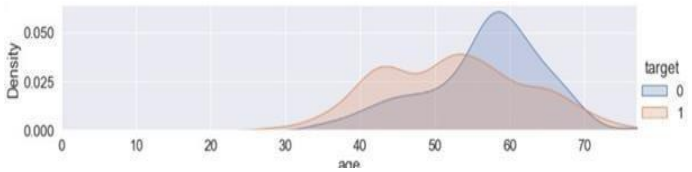



Facetgrid




Axisgrid






**ML Algorithm used KNN:**

KNN Among machine learning algorithms, neighbour is among the simplest. This method relies on the popular Supervised Learning framework. The K-Nearest Neighbor (KNN) algorithm is an example-based learning method. If this is the case, it indicates the system will not retain any training information, data but it will memorize the training data. We also call it a lazy learner because it is very lazy to process the data or information. KNN is a non- parametric algorithm, it means that the KNN algorithm model does not make assumptions on underlying data KNN algorithm can be understood by calculating euclidean distancefor a given data query.

We select those nearest neighbors who are minimal and close to the value of 'K'. Therefore, it's employed most often as a sorting algorithm, K-means relies on the observation that points with similar characteristics tend to be located close to one another.

When solving classification issues, it is common practise to choose the label that is most often associated with a specific piece of data. Plurality voting is a more precise phrase for this situation, however "democratic decision" is more generally used in written works. By definition, "majority voting" requires more than half of voters in order to pass, and so is only applicable when only two options are on the table. You may give a strategically aligned with a score of larger than 25% once you have many classes, for example four classes, and you will not need 50percent of the majority to create a result of a pair ofclasses.

Implementation of KNN algorithm on our heart diseasedataset:

1. We will import all the libraries.
2. We will read the heart disease dataset.
3. Then we will perform KNN by splitting to train and test theset.
4. We will calculate for the best value of K.
5. And then we will apply the KNN Algorithm.
6. After all the above process we will test the accuracy. Acc- ording to the accuracy we will perform more hyperparametertuning for improvement.

**Logistic Regression:**

Just like KNN algorithm Logistic regression is also based on supervised learning technique. This statistical model (sometimes called a logit model) is often used in the fields of categorization and analytic applications. With a given collection of independent variables, such as whether or not a person votes, the likelihood of a certain event, such as whether or not a person votes, may be estimated using logistic regression. The top result might be a probability, therefore the range of the variable is between 0 and 1. The following formula may be thought of as a symbol for the logistic function:

$$Y = 1/1+e^{(-x)}$$

The logistic function, sometimes referred to as the sigmoid function, just transforms the different variables into a probability expression that has a range of 0 to 1 with regard to the dependent variable. Regardless of the prediction, logistic regression converts it into a probability that ranges from 0 to 1. Logistic regression is a member of the machine learning supervised model family in the context of artificial intelligence. Additionally, it is thought of as a due to the different types, indicating that it makes an effort to discern across classes (or categories).

It cannot, as the name indicates, build info, sort more or less an image, of a class that it's trying to forecast, unlike a generative method like naive Bayes[14] (e.g. a photo of
the cat). How logistic regression works:

The Logistic Regression algorithm works as follows –

Logistic regressive maps the real values of the independent variable between the interval of 0 and 1. The cutoff point is on
0.5. So, there will be the values which will be lying above 0.5 and there will be the values which will be lying below 0.5 cutoff. Therefore, logistic regression classifies the below cutoff values as class B which will indicate there is very low possibility of a certain occurrence and above cutoff values as class A which will indicate that

there is a high possibility of certain occurrence.

**Implement Linear Equation**

In order to calculate a response value, the logistic regression procedure employs a straight-line equation with explanatory or independent factors. We take the example of the amount of hours examined and the likelihood of passing the test, for instance. Here, the explanatory variable is the number of study hours, and it is represented by the symbol x1.

The response or goal variable represents the chance ofpassing the test, and it is represented by the letter z.

The following equation is the linear equation if we haveone parameter (x1) and one predict-and (z).

$z = \beta 0 + \beta 1 x1$

Here, the values with β0 and β1 are the parameters of themodel.

If there are different self telling variables, then the equation mentioned above could be expanded to $z = \beta 0 + \beta 1 x 1 + \beta 2 x 2 + \beta n x n$

Here, the values with β 0, β 1, β 2, and β n are the limitationsof proposed model.

**Sigmoid Function**

Next, a probability value between 0 and 1 is converted from the predicted response value, denoted by the symbol z. We use the sigmoid function to translate anticipated values into probability values. This sigmoid function then converts anyreal integer into a chance value ranging from zero and one.

Machine learning transforms predictions into probabilities using the sigmoid function. The sigmoid function shows anS-shaped curve. It is also known as a sigmoid arc.

The sigmoid function is a particular kind of logistic function. It may be obtained using the mathematical formula below.

The graph below can be used to visually depict the sigmoidfunction:

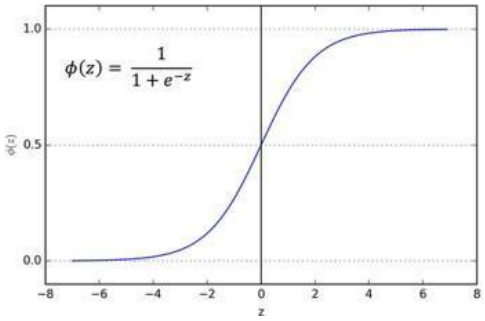



Figure 3.

**Decision Boundary**

The sigmoid function yields a probability value, which ranges from 0 to 1. A discrete class with the values "0" or "1" is then assigned to this probability value. We choose a threshold value in order to transform this statistical significance to a distinct class (pass/fail, yes/no, true/false). The Decision Boundary is the name given to this value as the threshold. Just below that threshold level, we will map data into class 0, and beyond this threshold level, we will apply values into class 1.

With Mathematical solution , it can be expressed as follows:-$p \geq 0.5 \Rightarrow$ class = 1

$p < 0.5 \Rightarrow$ class = 0

Basically, the decisive boundaries are set to be 0.5. So ,incase the chances of p value to be 0.8 (>0.5), we have to map thoseresults to class 1.

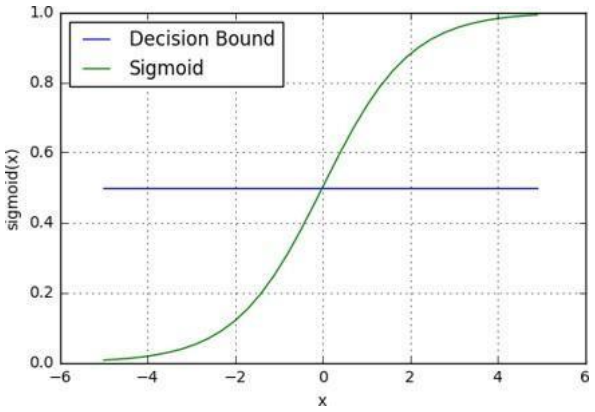



Figure 4.

Hence, if in this case, the chances of p value to be 0.2 (<0.5),we have to map those results to class 0.

**Random Forest:**

Using a Random Forest, a popular machine learning approach, that aggregates the outputs of various decision trees to produce a single result. Leo Breiman [12] and Adele Cutler[13] have registered the Random Forest trademark.

Since it can deal with both classification and regression issues, its adaptability and simplicity have been a driving force in its popularity.

To increase the dataset's predicted accuracy, a Random Forest uses using several decision trees on various slices of the supplied information and averaging the results of them. The more trees there are forest, the greater the precision and less over fitting there will be.

One method for facilitating choice is the decision tree. A graph or a simulation of choices and their possible results, including the value of resources, the price of those resources, and the results of random events. It's one example of how to demonstrate an algorithm. Machine learning decision trees are widely used in the field of operations research, particularly in the field of decision analysis, to determine the best course of action for accomplishing a given goal. Training models for predicting the class or quantity of both the target variable may be generated using a Decision Tree by learning simple decisionrules learned from past data (training data).

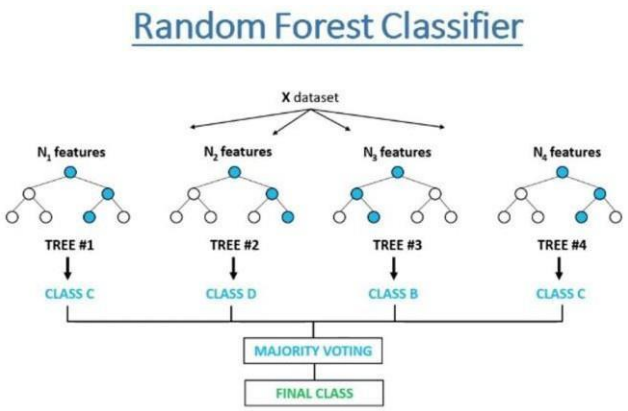


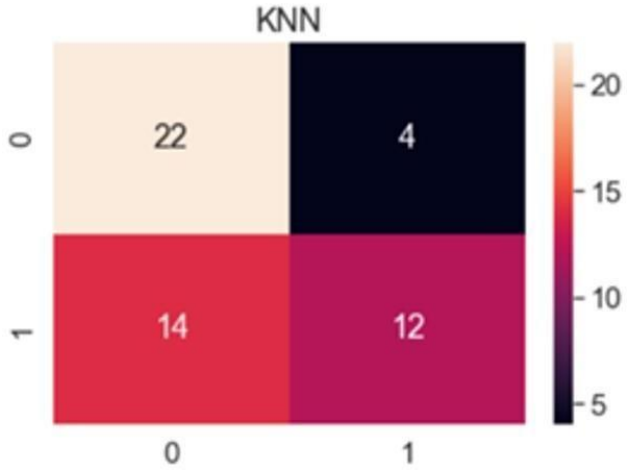

Figure 5.

**Comparison of Algorithm used**

Classification report for KNN

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.52 | 0.55 | 0.54 | 20 |
| 1 | 0.71 | 0.69 | 0.70 | 33 |

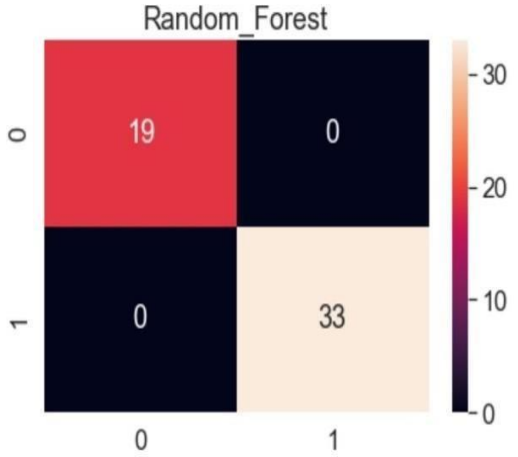
| | | | | |
|---|---|---|---|---|
| accuracy | | | 0.63 | 52 |
| macro avg | 0.62 | 0.62 | 0.62 | 52 |
| weighted avg | 0.64 | 0.63 | 0.64 | 52 |




Random Forest Classification Report

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 1.00 | 1.00 | 1.00 | 20 |
| 1 | 1.00 | 1.00 | 1.00 | 33 |

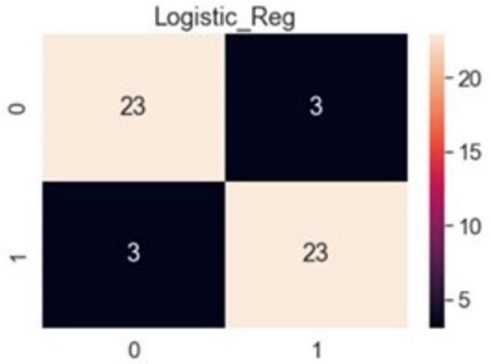
| | | | | |
|---|---|---|---|---|
| accuracy | | | 1.00 | 52 |
| macro avg | 1.00 | 1.00 | 1.00 | 52 |
| weighted avg | 1.00 | 1.00 | 1.00 | 52 |




Classification report for Logistic Regression

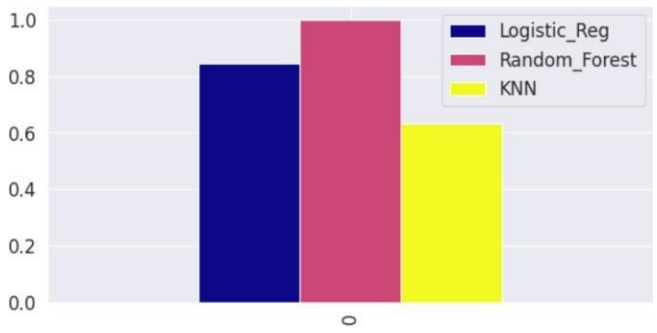
| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.80 | 0.80 | 0.80 | 20 |
| 1 | 0.88 | 0.88 | 0.88 | 33 |

| | | | | |
|---|---|---|---|---|
| accuracy | | | 0.85 | 52 |
| macro avg | 0.84 | 0.84 | 0.84 | 52 |
| weighted avg | 0.85 | 0.85 | 0.85 | 52 |




**Accuracy Comparison**

| Logistic Regression | Random Forest | KNN |
|---|---|---|
| 0.85 | 1.00 | 0.63 |



**Conclusion and Future Scope**

As we can see from the data above, Random forest, followed by Logistic Regression and KNN, is the most accurate Machine Learning method for predicting cardiovascular illness. As a result, we draw the conclusion that Random forest, out of the three algorithms, is the best at predicting heart disease in patients.

Numerous algorithms, including Inference using support vector machine ( svm and the Naive Bayes method, are included in the machine learningfield. By employing techniques like feature engineering and modifying hyperparameters, these algorithms can be made better.

As a result, we can conclude that in order to get even better results, the same problem should be tackled using differentML algorithms.

**References**


[1]. https://www.mayoclinic.org/diseases-conditions/heart- attack/symptoms-causes/syc-20373106
[2]. Animesh Hazra, Arkomita Mukherjee, Amit Gupta, Asmita Mukherjee, "Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review", Research Gate Publications, July 2017, pp.2137- 2159.
[3]. https://www.nhs.uk/conditions/cardiovascular-disease/
[4]. Bhardwaj, A., Kundra, A., Gandhi, B., Kumar, S., Rehalia, A. and Gupta, M., 2019. Prediction of heart attack using machine learning. IITM Journal of Management and IT,10(1), pp.20-24.
[5]. Chauhan, Y.J., 2018. Cardiovascular disease prediction using classification algorithms of machine learning. Int. J. Sci. Res. ISSN, pp.2319-7064.
[6]. Balakrishnan, M., Christopher, A.A., Ramprakash, P. and Logeswari, A., 2021, February. Prediction of Cardiovascular Disease using Machine Learning. In Journal of Physics: Conference Series (Vol. 1767, No. 1, p. 012013). IOP Publishing.
[7]. Prerana T H M1, Shivaprakash N C2 , Swetha N3 "Prediction of Heart Disease Using Machine Learning Algorithms- Naïve Bayes,Introduction to PAC Algorithm, Comparison of Algorithms and HDPS" International Journal of Science and Engineering Volume 3, Number 2 – 2015 PP:90-99
[8]. ©IJSE Available at www.ijse.org ISSN: 2347-2200
[9]. B.L Deekshatulua Priti Chandra "Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm " M.Akhil jabbar* International Conference on ComputationalIntelligence: Modeling Techniques and Applications (CIMTA) 2013.
[10]. Michael W.Berry et.al,"Lecture notes in datamining",World Scientific(2006).
[11]. S. Mohan, C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in IEEE Access, vol. 7, pp. 81542-81554, 2019,doi: 10.1109/ACCESS.2019.2923707.
[12]. Nikhar, S., & Karandikar, A. M. (2016). Prediction of heart disease using machine learning algorithms. International Journal of Advanced Engineering, Management and Science, 2(6), 239484.
[13]. Breiman, L. (2001). Random forests. Machine learning,45(1), 5-32.
[14]. Cutler, Adele, D. Richard Cutler, and John R. Stevens. "Random forests." Ensemble machine learning. Springer, Boston, MA, 2012. 157-175.
[15]. Huang, Y., & Li, L. (2011, September). Naive Bayes classification algorithm based on small sample set. In 2011 IEEE International conference on cloud computing and intelligence systems (pp. 34-39). IEEE.
[16]. S. Shilaskar and A. Ghatol, "Feature selection for medical diagnosis : Evaluation for cardiovascular diseases," Expert Syst. Appl., vol. 40, no. 10, pp. 4146–4153, Aug. 2013.
[17]. T.Mythili, Dev Mukherji, Nikita Padaila and Abhiram Naidu, "A Heart Disease Prediction Model using SVMDecision Trees-Logistic Regression (SDL)", International Journal of Computer Applications, vol. 68, 16April 2013.
[18]. Sharan Monica.L, Sathees Kumar.B, "Analysis of CardioVasular Disease Prediction using Data Mining Techniques", International Journal of Modern ComputerScience, vol.4, 1 February 2016, pp.55-58.


Figure [1]: Centers for Disease Control and PreventionFigure [2]: Mayo Clinic
Figure [3]: Towards Data Science
Figure [4]: https://medium.com/analytics-vidhya/decision- boundary-for-classifiers-an-introduction-cc67c6d3da0e
Figure [5]: freeCodeCamp