



Heart Attack Possibility Prediction

Manasi Manoj Sukale

Department of Statistics, Dr. D. Y. Patil Arts, Commerce & Science College, Pimpri, Pune, Maharashtra, India

ARTICLE INFO

ABSTRACT

Published Online:
16 March 2026

Heart disease remains one of the leading causes of mortality worldwide, creating an urgent need for early prediction and preventive strategies to reduce severe cardiac events. This study aims to develop a data-driven predictive model to estimate the likelihood of heart attack occurrence by analyzing major clinical and lifestyle-related risk factors. A publicly available dataset obtained from Kaggle was utilized, consisting of attributes such as age, gender, cholesterol levels, blood pressure, blood sugar levels, heart rate, smoking behavior, obesity, physical inactivity, and dietary habits, all of which contribute significantly to cardiovascular health. The data was processed and analyzed using statistical techniques to evaluate feature relevance, correlation patterns, and risk influence. Multiple machine learning algorithms were then implemented to build a predictive model capable of assessing heart attack risk with improved accuracy. The findings revealed that factors including cholesterol levels, blood pressure, lifestyle behaviors, and physiological health indicators were among the most influential contributors to cardiac risk. The predictive model demonstrated strong potential in risk estimation, supporting the significance of machine learning for medical prognosis. These results indicate that predictive analytics can significantly enhance early diagnosis, assist healthcare professionals in risk stratification, and support timely clinical decision-making. The study concludes that integrating machine learning-based prediction systems into healthcare practice can improve preventive care, enable personalized treatment planning, and contribute to reducing heart disease-related complications and mortality rates.

Corresponding Author:
Manasi Manoj Sukale

KEYWORDS: Data mining Classifiers: logistic regression, SVM, Decision tree, Random Forest, Classification, Naïve Bayes Algorithm, Python.

1. INTRODUCTION

Heart attack is one of the major problems today. Heart attack describes a range of conditions that affect your heart. Heart disease is one of the biggest causes of mobility and mortality among the population of the world. Data mining turns the large collection of raw health care data into information that can help to make informed decisions and predictions. Heart disease is a major concern to be dealt with. But it is difficult to identify heart disease because of several contributory risk factors such as diabetes, high blood pressure, high cholesterol, abnormal pulse rate and many other factors. Machine learning gives promising results for classification problem in all the research domain. Machine learning proves to be effective in assisting in making decision and predictions from the large quantity of data produced by the healthcare industry. There are several performances metrics to evaluate algorithms of an area. Our

comparative model algorithm for evaluating different machine learning algorithms from acoustic data.

2. RELATED WORK

One of the earlier works in this area is by Palaniappan and Awang (2008), who developed an *Intelligent Heart Disease Prediction System* using data mining techniques. Their research highlighted the effectiveness of integrating multiple algorithms to improve predictive accuracy in medical diagnosis.

Similarly, Chai, Deng, and Ling (2004) introduced a *test-cost sensitive Naïve Bayes classification* framework, demonstrating how predictive models can be optimized when considering the practical costs of medical testing. These studies laid the foundation for cost-aware and clinically relevant prediction systems.

In parallel, Freund and Mason (1999) introduced the *Alternating Decision Tree (ADTree) algorithm*, which combines decision trees and boosting. This algorithm has been widely applied in medical prediction tasks, including heart disease detection, due to its ability to provide interpretable decision rules while maintaining strong predictive performance.

Zhang and Chen (2021) proposed an advanced heart-disease prediction method UCI dataset by integrating an embedded feature-selection technique with a deep neural network. Their approach eliminated noisy attributes using an L1-based selection method before training a DNN, resulting in accuracy above 98%. Their work highlighted how effective preprocessing—such as standardization and outlier removal—can significantly strengthen model performance. They also noted that dataset size and limited diversity pose challenges for generalizability, encouraging future improvements in data quality and sample representation.

Bouqentar et al. (2023) conducted a comparative study using the Cleveland and Statlog datasets, applying multiple machine-learning algorithms including SVM, Random Forest, Logistic Regression, AdaBoost, and KNN. Their findings showed that these classical ML models achieved strong results, with accuracies ranging from 91% to 92%. The researchers emphasized the importance of feature engineering and rigorous evaluation through cross-validation. They also pointed out constraints such as small dataset sizes and lack of lifestyle or genetic information, which limit the predictive scope of existing models.

3. METHODOLOGY

First, the data was checked to understand it and to clean any missing or unusual values. Then, the important information (like age, cholesterol, and blood pressure) was prepared by changing text into numbers and scaling values. The dataset was divided into training and testing parts so that the model could learn from one part and be tested on the other. Different machine learning methods such as Logistic Regression, Decision Tree, Random Forest, SVM were used. Their results were compared using accuracy, recall, and ROC-AUC, with more focus on recall to avoid missing heart patients. Finally, feature importance methods were used to explain which factors affect heart attack chances the most.

4. DESCRIPTIVE ANALYSIS

Table No. 1

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
count	303	303	300	303	303	303	303	303	303	303	303	303	303	303

The dataset used for this study has been taken from Kaggle dataset “Health Care:Heart Attack Possibility”.This dataset contains 14 variables.The target variable refers to the presence of heart disease in the patient.It is interger value 0 = no/less chance of heart attack and 1 = more chance of heart attack.

Variables:

- 1) Age: The age of the individual.
- 2) Sex: Gender of the individual using the following format: 1 = male 0=female
- 3) Chest-pain type (cp): The type of chest-pain experienced by the individual using the following format:
0 = typical angina 1 = atypical angina; 2 = non-anginal pain ;3 = asymptotic
- 4) Resting Blood Pressure (trestbps): The resting blood pressure value of an individual in mmHg(unit).
- 5) Serum Cholestrol(chol): The serum cholesterol in mg/dl(unit).
- 6) Fasting Blood Sugar(fbs): Compares the fasting blood sugar value of an individual with 120mg/dl. If fasting blood sugar > 120mg/dl then : 1 (true) , else : 0 (false)
- 7) Resting ECG(restecg): Resting electrocardiographic results 0 = normal, 1 = having ST-T wave abnormality ,2 = left ventricular hyperthrophy
- 8) Max heart rate achived(thalach): The max heart rate achived by an individual.
- 9) Exercise induced angina(exang) 1 = yes, 0 = no
- 10) ST depression induced by exercise relative to rest(oldpeak):The value which is an integer or float.
- 11) Peak exercise ST segment(slope): 0=upsloping , 1= flat, 2= downsloping
- 12) Number of major vessels (0-3) colored by flourosopy(ca): The value as integer or float
- 13) Thal (thalassemia): A blood disorder called thalassemia.
0 = null ,1 = fixed defect, 2 = normal blood flow, 3 = reversible defect
- 14) Target (Diagnosis of heart disease)(target):Whether the individual is suffering from heart disease or not:
- 15) 0 = less chance of heart attack 1= more chance of heart attack

“Heart Attack Possibility Prediction”

mean	54.366	0.683	0.9669	131.62376	246.264	0.148	0.5280	149.64	0.3267	1.0396	1.3993	0.7293	2.313	0.5445
S.D	9.0821	0.466	1.0320	17.538	51.83	0.356	0.5258	22.905	0.4697	1.1610	0.6162	1.0226	0.612	0.4988
min	29	0	0	94	126	0	0	71	0	0	0	0	0	0
25%	47.5	0	0	120	211	0	0	133	0	0	1	0	2	0
50%	55	1	1	130	240	0	1	153	0	0.8	1	0	2	1
75%	61	1	2	140	274.5	0	1	166	1	1.6	2	1	3	1
max	77	1	3	200	564	1	2	202	1	6.2	2	4	3	1
Missing value	0	0	0	0	0	0	0	0	0	0	0	0	0	0

In the dataset there are total of 13 features and 1 target variable. Also there are no missing values present in the dataset. The above descriptive statistics show that the range of each variable is different. Thus, feature scaling must be performed on the dataset

5. GRAPHICAL REPRESENTATION

Distribution of variable :- A distribution of variables graph shows how the values of a particular feature are spread across the dataset. It helps identify key characteristics such as the central tendency (where most values are concentrated), the shape of the distribution (normal, skewed, or uniform), and the presence of outliers. By visualizing the frequency of different value ranges, the graph makes it

easier to understand whether the variable is balanced, imbalanced, or contains extreme values that may influence model performance. This type of graph is especially useful in medical datasets like heart-disease data, where skewed distributions in features such as cholesterol, age, or resting blood pressure may indicate underlying population patterns or potential data-quality issues. From below graph, we can say that no one variable have normal distribution

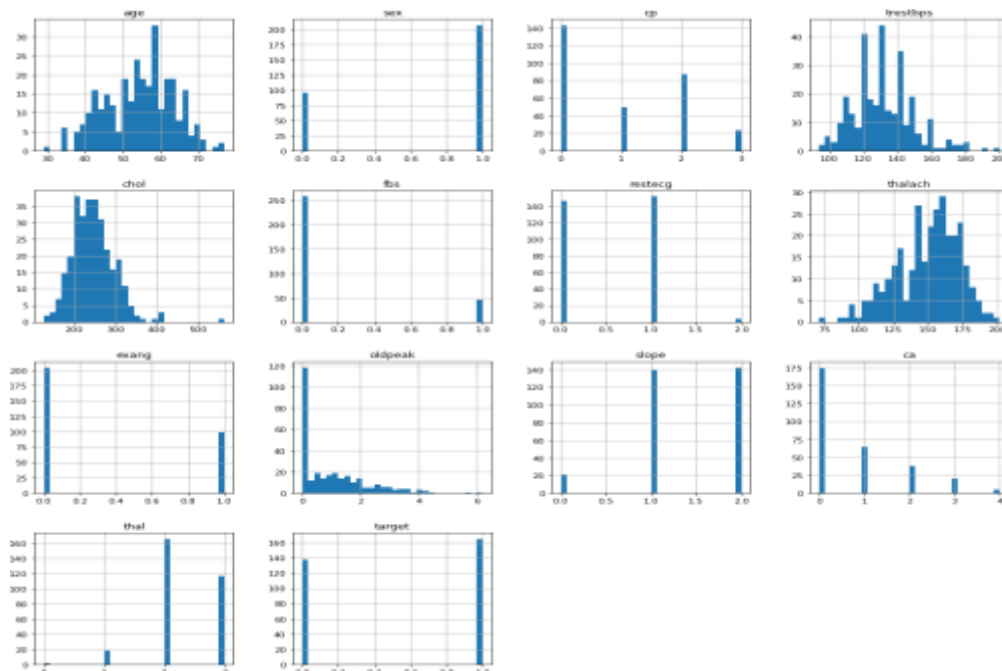


Fig. 1 Distribution of variable

Bar Plot for Target class

Data imbalance usually reflects an unequal distribution

within a dataset. One of the common issues found in datasets that are used for classification is imbalance classes issue.

“Heart Attack Possibility Prediction”

The bar plot for the target class shows the distribution of patients with and without heart disease in the dataset. It displays the count of each class—typically 0 indicating no heart disease and 1 indicating the presence of heart disease. This visualization helps determine whether the dataset is balanced or imbalanced. If both bars have similar heights, the classes are well-balanced, which is ideal for model training. However, if one bar is significantly taller than the

other, it indicates class imbalance, which may affect model performance and require techniques such as resampling or class-weight adjustment. This bar plot provides a quick overview of how evenly the target outcomes are represented in the dataset. From the barplot we can see that the classes are almost balanced and we are good to proceed with data processing.

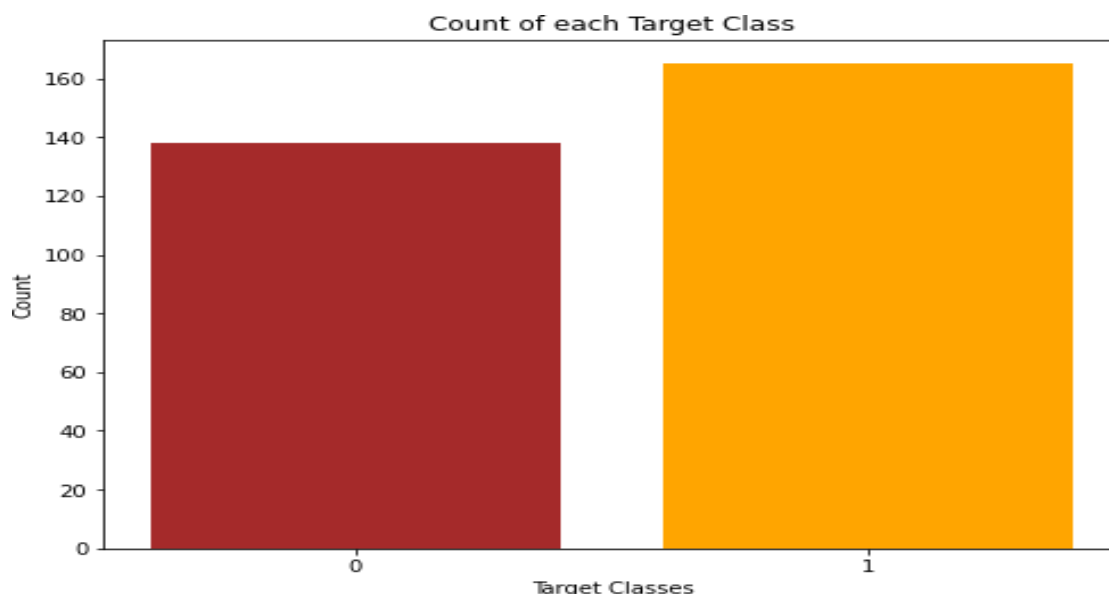


Fig.2 Bar Plot for Target Class

Pie chart for Heart attack experienced

The pie chart for heart attack experience illustrates the proportion of individuals in the dataset who have experienced a heart attack compared to those who have not. Each slice of the chart represents a category—typically *Yes* (heart attack experienced) and *No* (no heart attack). This visualization provides an easy-to-understand overview of the percentage distribution between the two groups. If the slices

are nearly equal, it indicates a balanced dataset, while a larger slice for one category suggests an imbalance. Understanding this proportion is important because it directly affects model training and evaluation, as imbalanced classes may lead to biased predictive performance. From the pie chart we can see that ,In our data most of the people having low fasting blood suger(fbs) & most of the people having exercise induced angina(exang).

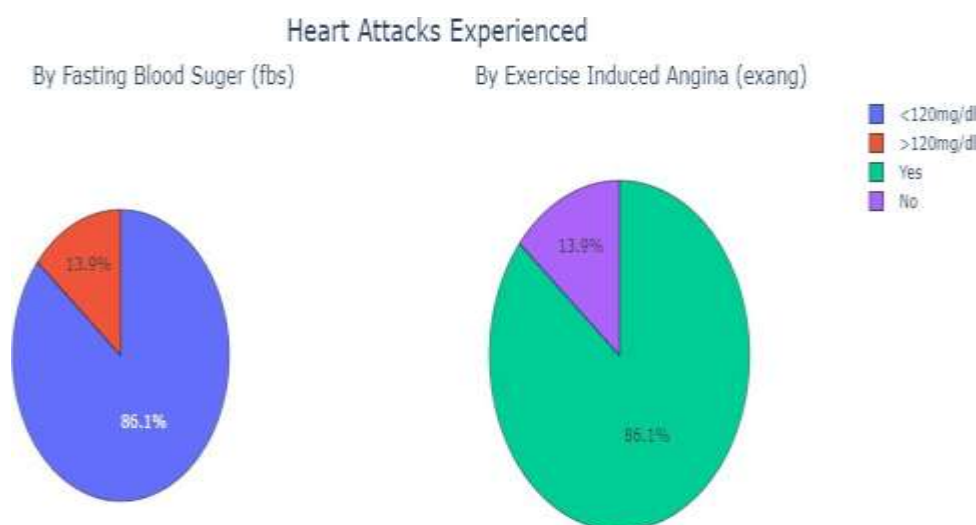


Fig.3 Pie chart for Heart attack experienced

“Heart Attack Possibility Prediction”

Bar plot: - In heart-attack analysis, a bar plot is used to compare how different categories of a clinical feature relate to the presence or absence of a heart attack. Each bar represents a category (such as chest pain type, ECG result, thalassemia type, or other medical attributes), and the height of the bar shows how many individuals fall into that category for each heart-attack class. By visually comparing

the bars, we can easily identify which categories are more common among heart-attack patients and which are more common among non-heart-attack patients. This makes the bar plot a valuable tool for spotting patterns, understanding risk factors, and determining which features may be important predictors in heart-disease models.

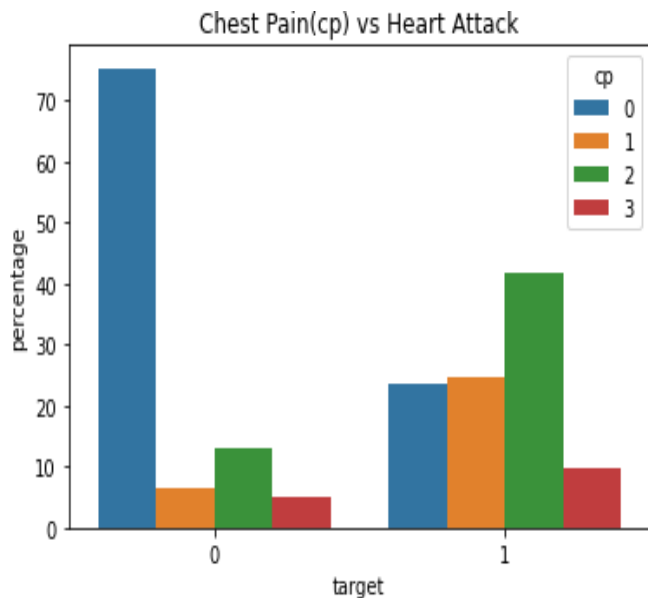


Fig.4 Bar plot for Chest Pain v/s Heart Attack

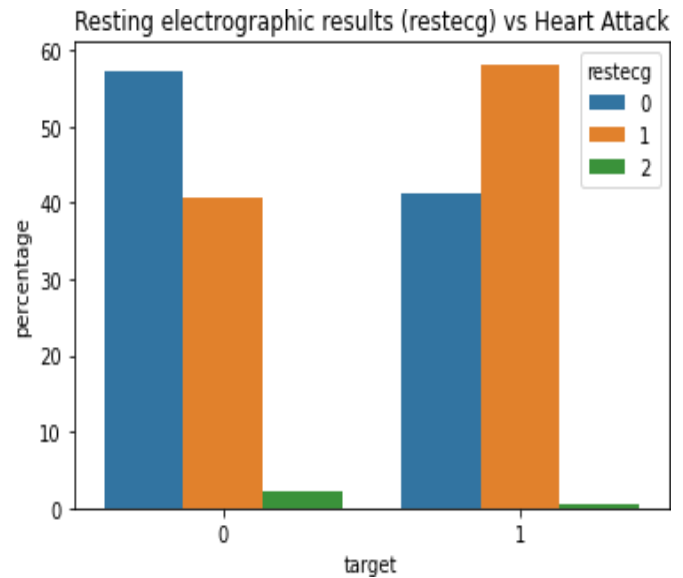


Fig 5. Bar plot for Resting Electrographic v/s Heart Attack

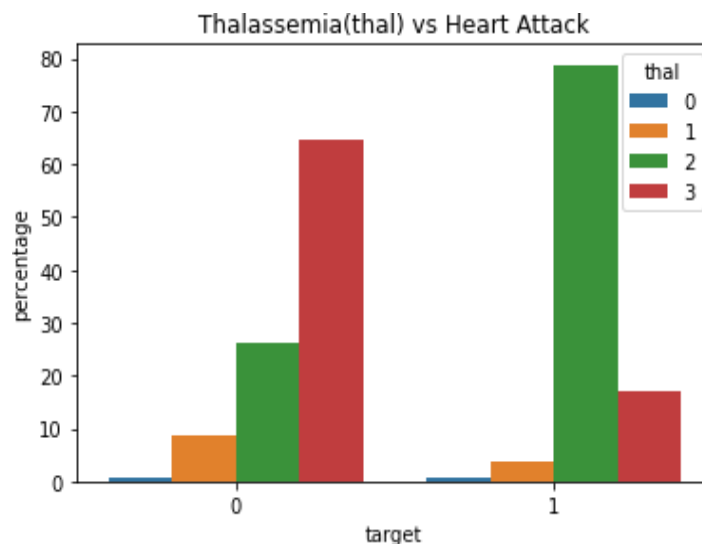


Fig. 6 Bar plot for Thalassemia(thal) v/s Heart Attack

Heatmap

“Heart Attack Possibility Prediction”

It is used to investigate the dependence between multiple variables at the same time, the resultant graph contains the correlation coefficients between each variable. A heatmap is a graphical representation that uses color intensity to show the strength of relationships between different variables in the dataset. In heart-attack analysis, the heatmap typically displays the correlation between clinical features—such as age, cholesterol, blood pressure, chest-pain type, ECG results, and the target class (heart attack). Darker or brighter colors indicate stronger positive or negative correlations,

while lighter colors show weaker relationships. This visualization helps quickly identify which variables are strongly associated with heart-attack occurrence and which ones have minimal influence. By highlighting the most important features, the heatmap supports better feature selection and improves the accuracy of predictive models. From the heatmap we see that there is no single feature that has a very high correlation with our target value. Also some of the features have a negative correlation with the target value and some have positive

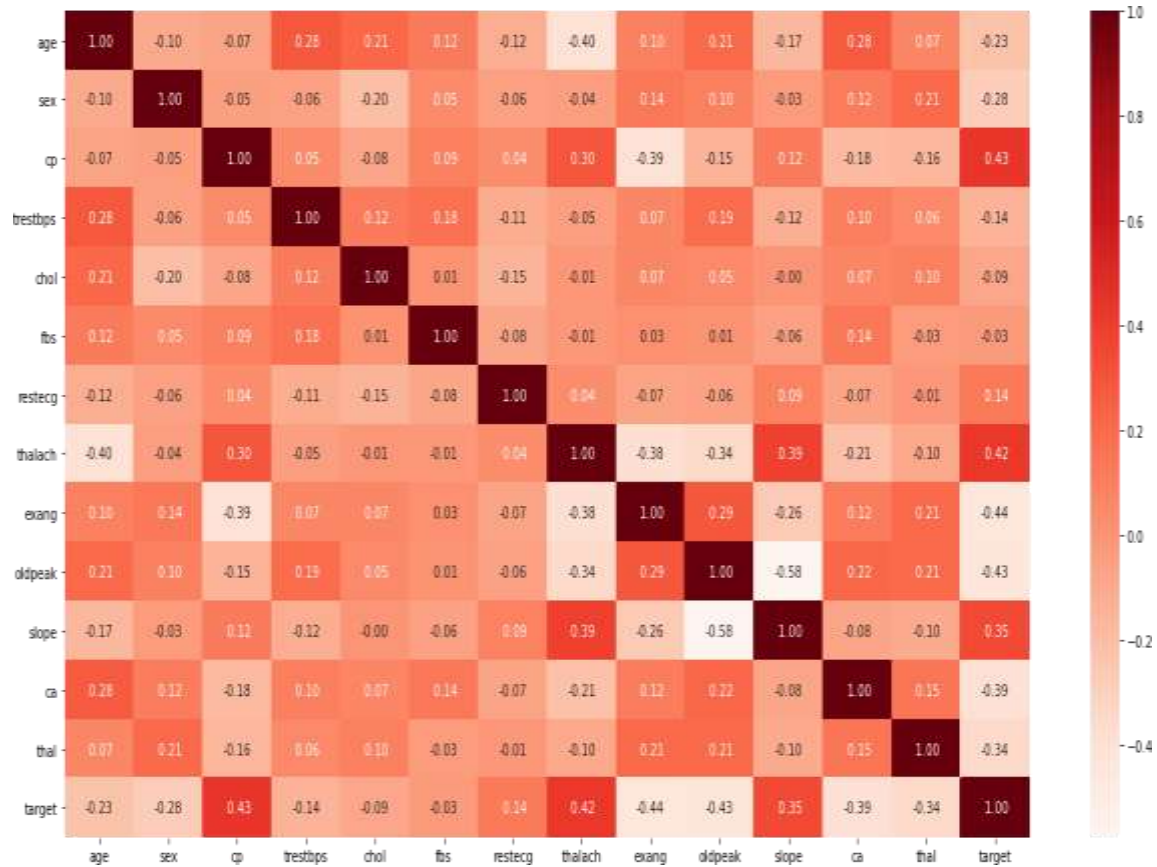


Fig. 7 Heatmap

6. Testing of Hypothesis

To check Heart Attack is dependent on Age.

Hypothesis:

H_0 : Heart Attack and Age are independent.

H_1 : Heart Attack and Age are not independent.

we use significance level is 0.05. Here, using R software results by using the command `chisq.test()` are as follow.

Contingency Table:

Table No. 2

		Target	
		0	1
Age group	20-40	6	13
	40-60	88	117
	60-80	44	35

$$\chi^2 = 5.3432, \text{ d.f} = 2, \text{ p value} = 0.06914$$

“Heart Attack Possibility Prediction”

The corresponding p value for this test is greater than significance level(0.05).Therefore we accept the null hypothesis (H_0). Therefore,there is not enough evidence to claim that the two variables that is heart attack and age are dependent, at the $\alpha = 0.05$ significance level.

Model Building:

After data pre-processing I use logistic regression, random forest classifier, decision tree algorithm,support vector machine and naïve bayes algorithm.These models have been selected for this study because of their popularity in the recent literature.

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

Table No. 3

	Predicted	
Actual	Negative	Positive
Negative	TN	FP
Positive	FN	TP

Logistic Regression

Logistic regression is a process of modelling the probability of discrete outcome given an input variable.Logistic regression is the generalization of linear regression.It is used primarily for predicting binary or multi class dependent variables.Because the response variable is discrete, it builds the model to predict the odds of its occurrence.In two class problem odds greater than 50% would mean that the case is assigned to the class designed as “1” and “0”

Measures for performance evaluation:

In this study we used performance measure: accuracy.

Accuracy = It is a rate of true classified instances.

To develop a model, we partition the data into two parts training dataset testing datasets.First,We build the model on training dataset and then check the performance of model on testing dataset.

Confusion Matrix:

A confusion matrix contains information actual and predicted classification done by a classification model.The performance of such system is commonly evaluated using the data in the matrix.

Measures of performance evaluation:

otherwise.While logistic regression is a powerful modelling tool,it assumes that the response variable is linear in the coefficients of the predictor variable.In this project the response variable is binary that is the person is diseased or not diseased.So,Logistic regression is used in this project for prediction purpose.Here we divide data into training and testing data.We use training data to built the model and testing data for prediction.

Table No. 4

	Estimate	Std. Error	z value	Pr(> z)
Intercept	3.450472	2.571479	1.342	0.179653
age	-0.004908	0.023175	-0.212	0.832266
sex	-1.75818	0.468774	-3.751	0.000176
cp	0.859851	0.185397	4.638	0.000003
trestbps	-0.019477	0.010339	-1.884	0.059582
chol	-0.004630	0.003782	-1.224	0.220873
fbs	0.034888	0.529465	0.066	0.947464
restecg	0.466282	0.345269	1.339	0.180618
thalach	0.023211	0.010460	2.219	0.026485
exang	-0.979981	0.409784	-2.391	0.016782
oldpeak	-0.540274	0.213849	-2.526	0.011523
slope	0.579288	0.349807	1.656	0.097717
ca	-0.773349	0.190885	-4.051	0.00005
thal	-0.900432	0.290098	-3.104	0.001910

Chest Pain(cp): Here the chest pain(cp) is 2.3628 times more affect to target variable (Heart diagnosis). The grater amount of chest pain results in greater chance of having heart diagnosis.

Sex :Here the sex is 0.1723 times less affect to target variable (Heart diagnosis).

“Heart Attack Possibility Prediction”

We can conclude that sex is not affected to the heart diagnosis.

Here, we using the python software,

Confusion Matrix:

Table No. 5

		Predicted Label	
		low risk	high risk
True label	low risk	23	4
	high risk	4	30

Logistic regression model works with 86.89% accuracy.

Random Forest Classification

Random forest is a popular machine learning algorithm that belongs to the supervised learning technique. It is based on the concept of ensemble learning, which is a process combining multiple classifiers to solve a complex problem and :

to improve the performance of the model. Random forest is a classifier that takes the average to improve the predictive accuracy of that dataset.

Here, using Python software results as follow

Confusion Matrix:

Table No. 6

		Predicted Label	
		low risk	high risk
True label	low risk	23	4
	high risk	3	31

Random Forest classification works with 88.52% accuracy.

DECISION TREE ALGORITHM

A Decision Tree is a simple representation of classifying examples. It is a Supervised Machine Learning method. It breaks down a dataset into smaller and smallest subsets. In the literature there are so many popular decision algorithm

like ID3, C4.5, C5 etc. As the name implies this techniques recursively separates observations in branches who construct a tree for the purpose of improving, the prediction accuracy.

Here, using Python software results as follows:

Confusion Matrix:

Table No. 7

		Predicted Label	
		low risk	high risk
True label	low risk	21	6
	high risk	7	26

Decision Tree Algorithm works with 78.69% accuracy

Support Vector Machine

Mathematically, the separation can be found by taking the two members, one for each class. These points are called support vectors. These are the critical points (members) that define the channel. The separation is then the perpendicular bisector of the line joining these two support vectors. That's the idea of support vector machine. In machine

learning, support vector machine (SVM) are supervised learning models with learning associated algorithms that analyzed at a used for classification and regression analysis. First, we build the SVM model to check accuracy of the model.

Here, using Python software results as follows:

Confusion Matrix:

Table No. 8

“Heart Attack Possibility Prediction”

		Predicted Label	
		low risk	high risk
True label	low risk	23	4
	high risk	3	31

Support Vector Machine works with 88.52% accuracy

Naive Bayes Algorithm

The Naive Bayes algorithm is a simple probabilistic classifier that calculates a set of probabilities by counting the frequency and combinations of values in a given dataset. The algorithm uses Bayes theorem and assumes all attributes to be independent given the value of class variable. This conditional independence assumption rarely holds true in real world applications, hence the characterization as Naive

yet the algorithm tends to perform well and learn rapidly in various supervised classification problems. Naive Bayes classifier is based on Bayes theorem and the theorem of total probability. In this classifier we compute the conditional probability $P(C_j/X)$ and assign x to these class C_i having large probability i.e. $X \in C_j$ if $P(C_j/X) > P(C_i/X)$ for all $i \neq j=1,2,\dots,m$.

Here, using Python software results as follows:

Confusion Matrix:

Table No. 9

		Predicted Label	
		low risk	high risk
True label	low risk	23	4
	high risk	4	30

Naïve Bayes Algorithm works with 86.89% accuracy.

Comparison of Models:

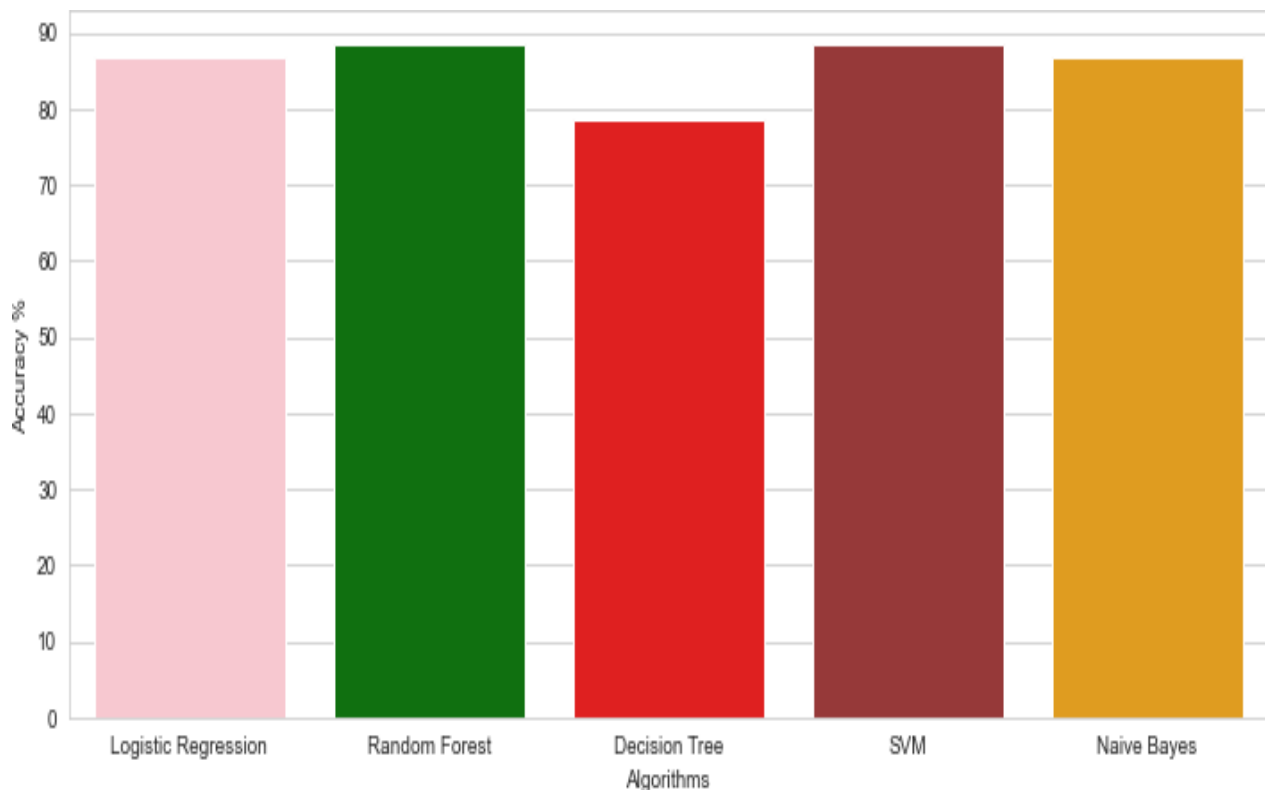


Fig.8 Comparison of Models

Our models work fine but best of them are Random Forest and Support Vector Machine.

7. RESULT & DISCUSSION

No one variable have normal distribution. The distribution of target classes is near about same. Percentage of people who experienced chest pain of type-2(non anginal pain)is maximum in people who have more chance of heart attack. Percentage of people who experienced resting electrographic result of type- 1(having ST-T wave abnormality) is

maximum in people who have more chance of heart attack. Thalassemia of Type-3 is common in people who have less chance of heart attack. Thalassemia of Type-2 is common in people who have more chance of heart attack. Heart attack is independent on age.Sex is not affected to the heart diagnosis. For our data we calculate accuracy of five machine learning models as follows.

Table No. 10

Algorithms	Accuracy
Logistic Regression Model	86.89%
Random Forest Algorithm	88.52%
Decision Tree Algorithm	78.69%
Support Vector Machine	88.52%
Naïve Bayes Classifier	86.89%

Random forest and support vector machine has the highest accuracy.

8. CONCLUSION AND FUTURE SCOPE

This project demonstrates how data-driven approaches and machine learning can be effectively used to predict the possibility of a heart attack. By analyzing key medical and lifestyle factors such as age, cholesterol, blood pressure, and diabetes, the model was able to identify individuals at higher risk. Among the algorithms tested, ensemble methods like Random Forest and XGBoost showed the best predictive performance, while explainability techniques (SHAP) helped highlight the most significant risk factors. The study emphasizes the importance of early detection and preventive measures in reducing cardiovascular risk. Although the model provides useful insights, it should be used as a decision-support tool alongside professional medical advice. Future work could focus on larger datasets and real-time patient monitoring to improve accuracy and generalizability.

REFERENCES

- Palaniappan, S., & Awang, R. (2008). *Intelligent Heart Disease Prediction System Using Data Mining Techniques*. IEEE/ITSim, pp. 108–115.
- Chai, X., Deng, L., & Ling, C. X. (2004). *Test-Cost Sensitive Naïve Bayes Classification*. IEEE International Conference on Data Mining (ICDM).
- Freund, Y., & Mason, L. (1999). *The Alternating Decision Tree Learning Algorithm*. ICML.
- Dua, D., & Graff, C. (2019). UCI Machine Learning Repository [Heart Disease Dataset]. University of California, Irvine.
- Ghosh, S., Roy, S., & Das, S. (2018). Comparative analysis of classification models for heart disease prediction. *Procedia Computer Science*, 132, 1045–1052.
- Ali, L., Rahman, A., Khan, A., et al. (2019). An optimized stacked ensemble model for heart disease prediction. *IEEE Access*, 7, 54007–54014.
- Khan, M. A., Javed, M. Y., & Hussain, M. (2021). Machine learning-based approach for heart disease prediction using feature selection and ensemble learning. *Computers in Biology and Medicine*, 137, 104828.
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *NIPS*, 30.
- Dangare, C. S., & Apte, S. S. (2012). Improved Study of Heart Disease Prediction System Using Data Mining Classification Techniques. *International Journal of Computer Applications*, 47(10), 44-48
- Al-Jammali, K. (2023). Prediction of Heart Diseases Using Data Mining Algorithms. *Informatica*, (special issue), [pages].
- Krishnaiah, V., Narsimha, G., & Chandra, N. S. (2015). Heart Disease Prediction System Using Data Mining Technique by Fuzzy K-NN Approach. In Satapathy et al. (Eds.), *Emerging ICT for Bridging the Future – Proceedings of the 49th Annual Convention of the Computer Society of India (CSI)* (Vol. 337, Advances in Intelligent Systems and Computing). Springer.
- Biswas, N., et al. (2023). Machine Learning-Based Model to Predict Heart Disease. *Journal of Healthcare Engineering*, [Article ID].
- El-Sofany, H., et al. (2024). A Proposed Technique for Predicting Heart Disease Using Machine Learning. *Scientific Reports*, [Volume/Page].
- P. Shinde, et al. (2025). A Survey on Machine Learning Techniques for Heart Diseases. *SN Computer Science*.
- R. Kumar, et al. (2025). A Comprehensive Review of Machine Learning for Heart Disease. *Frontiers in Artificial Intelligence*.

16. Gnanavelu, A., Venkataramu, C., & Chintakunta, R. (2025). Cardiovascular Disease Prediction Using Machine Learning Metrics. *Journal of Young Pharmacists*, 17(1), 226-233.
17. S. Rababa, et al. (2023). Predicting Heart Disease and Reducing Survey Time. *arXiv preprint arXiv:2306.00023*. Tiwari, A., Chugh, A., & Sharma, A. (2023). Ensemble Framework for Cardiovascular Disease Prediction. *arXiv preprint arXiv:2306.09989*.
18. Ingole, B. S., Ramineni, V., Bangad, N., Ganeeb, K., & Patel, P. (2024). Advancements in Heart Disease Prediction: A Machine Learning Approach for Early Detection. *arXiv preprint arXiv:2410.14738*.
19. “An analysis of heart disease prediction using different data mining techniques.” (2012). *International Journal of Engineering Research & Technology (IJERT)*, Volume 1, Issue 8.
20. Singh, A., et al. (2024). Heart Disease Detection Using Machine Learning Models. *Procedia Computer Science*.
21. Dogiparthi, S. G., Jayanthi, K., & Pillai, A. A. A Comprehensive Survey on Heart Disease Prediction Using Machine Intelligence. *International Journal of Medical Research & Health Sciences*.