# Heart disease prediction using machine learning algorithms

**AHEMRI Mariem**

Higher School of Technologies Essaouira (ESTE),Cadi Ayaad ,Marrakech ,Morocco

**Abstract:** Acute myocardial infarction (heart attack) is one of the deadliest diseases patients face. The key to cardiovascular disease management is to evaluate large scores of datasets, compare and mine for information that can be used to predict, prevent, manage and treat chronic diseases such as heart attacks. This pepper discusses the usage of machine learning in heart attack prediction , dataset used contained raw data of more than 300 file to train and test four machine algorithms: k-nearest neighbours, Random Forest, Regression Logistic and , decision tree . The performance of each algorithm in classifying heart attack is based on the accuracy and the Recall.

**Keywords** :Heart attack prediction , Machine learning, Random Forest, Logistic regression, decision tree, KNN

## I-Introduction

Heart, the main organ of the human body used for pumping blood into the whole body through the vessels of the circulatory system. In the circulatory system, the most crucial role is played by the heart . The circulatory system is the most important part of our body as it is responsible for the transport of blood carrying food, oxygen, water, minerals, and other important substance important for our body throughout our entire body. If the working of the heart is disrupted due to any circumstance and it does not function properly then it may cause serious health issues including death. Cardiovascular is a term that is used to refer to the pathologies altering or affecting the function or structure of the heart or blood vessel having the most common type of cardiovascular disease as coronary artery disease. The prevalence of the most common cardiovascular diseases (CVDs) represents the pinnacle of incurable processes which involve complex interactions between risk factors which can and cannot be improved. Most of the instances or cases of cardiovascular diseases can be allocated to improvable risk factors where most of the instances are considered preventable.

Major symptoms of the occurrence of heart attack are tightfistedness or affliction majorly in the chest, neck, back, and arms, tiredness, dizziness, abnormal heartbeat, and consternation. Risk factors including unchangeable factors like age, sex, family background, and changeable factors like smoking, high cholesterol, high blood pressure, fatness, deficiency of proper diet as well as exercise, and a huge amount of stress. Arterial reclamation to medication, ECG, and bypass surgery is the most used treatment methods in case of heart attacks .

## II-RELATED WORKS

As per initial prospects, ML can enhance the evolution of predictive models in health care industries, it was decided to test various algorithms to check what extent their prediction

scores estimate or ameliorate upon the results acquired in the authentic Framingham model. This model is one of the most important cardiac arrest risk prediction tables from the point of view of clinical practice. Several CV-risk prediction algorithms have been proposed till now [1]. The performance of various algorithms that have been used for calculating CV-risk can still be considered as a problem. Particularly, different scores like, Framingham Score, Systematic Coronary Risk Evaluation tend to underrate cardiac risk in patients [2]. Comparative study has been done for CV risk prediction using some well-known ML algorithms like k-nearest neighbours, support vector machine, classification, gradient boosting, logistic regression and regression tree and random forest[3]. Among the comparison between various features selection ML algorithms, for heart attack prediction Naive Bayes, SVM, and KNN the most optimistic classifiers.

There were two datasets Cleveland and Framingham which received classification accuracies of 93% and 91% respectively [4]. A research was carried out during July 2020 to make a model which predict the risk of cardiovascular disease using two different techniques. The support vector machine(SVM) was first trained and tuned perfectly for its parameters and then on training the SVM model for 1000 times , the average accuracy attained for the model to predict the cardio-vascular disease accurately was up to 96.5% with its average recall rate 89.8% while the recall rate using K-nearest neighbours reaches to 92.9% [5].

## III-Dataset and pre-processing:

### a)Dataset

The used dataset is a raw data of a group of patients contained 14 features listed below:

**age** - Age of the patient ,

**sex** - Sex of the patient ,

**cp** - Chest pain type ~ 0 = Typical Angina, 1 = Atypical Angina, 2 = Non-anginal Pain, 3 = Asymptomatic ,

**trtbps** - Resting blood pressure (in mm Hg) ,

**chol** - Cholesterol in mg/dl fetched via BMI sensor ,

**fbs** - (fasting blood sugar > 120 mg/dl) ~ 1 = True, 0 = False

**restecg** - Resting electrocardiographic results ~ 0 = Normal, 1 = ST-T wave normality, 2 = Left ventricular hypertrophy

**thalachh** - Maximum heart rate achieved

**oldpeak** - Previous peak

**slp** - Slope

**caa** - Number of major vessels

**thall** - Thalium Stress Test result ~ (0,3)

**exng** - Exercise induced angina ~ 1 = Yes, 0 = No

**output** - Target variable

### b)pre-processing:

After applying EDA obviously there are no NaN values in the data.There are certain outliers in all the continuous features.The data consists of more than twice the number of people with `sex` = 1 than `sex` = 0.There is no apparent linear correlation between continuous variable according to the heatmap.The scatterplot heatmap matrix suggests that there might be some correlation between `output` and `cp`, `thalachh` and `slp`.It is intuitive that elder people might have higher chances of heart attack but according to the distribution plot of `age` wrt `output`, it is evident that this isn't the case.According to the distribution plot of `thalachh` wrt `output`, people with higher maximum heart rate achieved have higher chances of heart attack.According to the distribution plot of `oldpeak` wrt `output`, people with lower pevious peak achieved have higher chances of heart attack.

## IV-Approach Performance

As mentioned, the performance of each algorithm in classifying heart attack prediction is based on the accuracy, detection rate.

The accuracy and detection rate are calculated using the confusion matrix. Confusion matrix, also called error matrix, is a table that shows different predictions and test results and compares them with the actual values. These matrices are used in statistics, data mining, machine learning models and other artificial intelligence applications.

| | | Predicted | |
|---|---|---|---|
| | | Negative (**N**) - | Positive (**P**) + |
| Actual | Negative - | True Negatives (T**N**) | False Positives (F**P**) **Type I error** |
| | Positive + | False Negatives (F**N**) **Type II error** | True Positives (T**P**) |

*Figure 1: Confusion Matrix*

**Accuracy:** Accuracy is the simplest. It defines your total number of true predictions in total dataset. It is represented by the equation of true positive and true negative examples divided by true positive, false positive, true negative and false negative examples.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

**Detection Rate(DR) :** indicates the ratio of the number of instances correctly classified as attack to the total number of attack instances present in test set.

$$DR = \frac{TP}{TP+FN}$$

Algorithms Performance The following table shows the results of the performance criteria values for each of the tested algorithms:

| Algorithms | Accuracy % | Recall % |
|---|---|---|
| Logistic Regression | 90.16 | 87.5 |
| Decision Tree | 78.68 | 71.87 |
| Random Forest | 78.68 | 71.87 |
| KNN | 86.88 | 87.5 |

## V-Discussion

From the experimental results , we can notice that the logistic Regression algorithm achieved the highest accuracy and the best in terms of in detection rate, on other hand the Random Forest and Decision Tree achieved the same results.

## VI-Conclusion

In this research, various Supervised ML classifiers namely, Random Forest, Decision Tree, KNN, and Logistic Regression have been used to deploy a model for Heart disease prediction. All the models achieved above 78% of accuracy. Suraj Kumar Gupta, Aditya Shrivastava, S. P. Upadhyay, Pawan Kumar Chaurasia (2021) used two datasets and attain an average accuracy equal to 85.5% and a recall rate of 82% in Framingham dataset based on Logistic regression classifier,

The results of the models were quit unsatisfying due to small dataset used in this research.

## References


1. J. J. Beunza et al., “Comparison of machine learning algorithms for clinical event prediction (risk of coronary heart disease),” J. Biomed. Inform., vol. 97, p. 103257, Sep. 2019, doi: 10.1016/J.JBI.2019.103257.
2. L. Navarini et al., “Cardiovascular Risk Prediction in Ankylosing Spondylitis: From Traditional Scores to Machine Learning Assessment,” Rheumatol. Ther., vol. 7, no. 4, pp. 867–882, 2020, doi: 10.1007/s40744-020-00233-4.
3. I. D. Mienye, Y. Sun, and Z. Wang, “An improved ensemble learning approach for the prediction of heart disease risk,” Informatics Med. Unlocked, vol. 20, p. 100402, Jan. 2020, doi: 10.1016/J.IMU.2020.100402.
4. M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, “Disease Prediction by Machine Learning over Big Data from Healthcare Communities,” IEEE Access, vol. 5, pp. 8869–8879, 2017, doi: 10.1109/ACCESS.2017.2694446.
5. P. Kaur, M. Sharma, and M. Mittal, “Big Data and Machine Learning Based Secure Healthcare Framework,” Procedia Comput. Sci., vol. 132, pp. 1049–1059, Jan. 2018, doi: 10.1016/J.PROCS.2018.05.020.