# Research Article Feb 2021

# Heart disease prediction using machine learning algorithms

Awais Ahmad

School of System and Technology

University of Management and Technology

Lahore, Pakistan

F2017065074@umt.edu.pk

Amanullah Nasir

School of System and Technology

University of Management and Technology

Lahore, Pakistan

F2017065056@umt.edu.pk

Muhammad Burhan

School of System and Technology

University of Management and Technology

Lahore, Pakistan

F2017065054@umt.edu.p

**Abstract** Heart illness is one reason for death these days. Foreseeing cardio ailment remains an intense test in the area of clinical information examination. Machine learning should be approved and to assist with choosing picking and hopes commencing the vast volume of measurements made by the clinical administration industry. We have in a like manner seen ML methodologies being used in progressing headways in

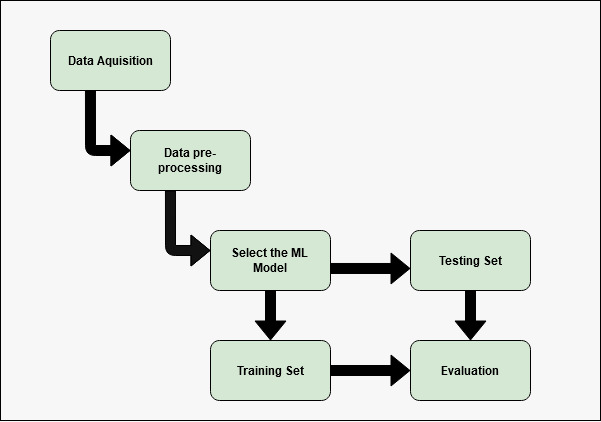
different locales of the Internet of Things. Various examinations give simply a short investigation foreseeing coronary sickness with ML techniques. Our concern of this research is to design a novel system that objectives finding enormous features by applying ML techniques achieving improving the precision in the conjecture of cardio ailment. The estimated model is given different blends of features and a couple alluded to game plan systems, for instance, Logistic Regression, Naive Bayes, K-Nearest Neighbors, Decision tree, and Random Forest

**Keywords**: machine learning, heart disease prediction, machine learning techniques, Logistic Regression, Naive Bayes, K-Nearest Neighbors, Random Forest

# Introduction

Utilizing Machine learning approaches has unbelievable potential for examining the concealed plans in the enlightening files of the clinical space. These models can be utilized for clinical ends. In any case, the open unrefined clinical data is extensively appropriated, heterogeneous in nature, and voluminous. This data ought to be accumulated in an organized construction. This accumulated data can be then fused to shape a crisis facility information system. Machine learning development gives a customer-organized methodology to managing and covered plans in the statistics. The major objective of our paper is to get comfortable with the exceptional methods of Machine learning used in the forecast of heart sickness. Existence is reliant on the compelling functioning of the heart because the heart is a major piece of our body.[[1](#_ENREF_1)] In case the movement of the heart isn't suitable, it will impact the other body segments of individuals, for instance, the cerebrum, kidney, etc heart disease is contamination that impacts the action of the heart. Different factors grow the peril of heart affliction. Nowadays, on the planet Heart sickness, is the critical explanation behind passing. The World Health Organization (WHO) addresses that 12 million deaths happen around the globe, reliably because of Heart disorders. In 2008, 17.3 million people passed on because of Heart Sickness. Over 80% of passing on the planet are a consequence of Heart ailment. WHO evaluated by 2030, basically 23.6 million people will kick the basin due to Heart disease as written in. Gauge by using data mining techniques gives us accurate eventual outcomes of the ailment. The heart sickness assumption structure can discover and focus on covered data identified with heart disease from a recorded coronary ailment informational index. It can answer complex requests for diagnosing coronary ailment and along these lines help clinical consideration agents and specialists to settle on smart clinical decisions that standard decision genuinely steady organizations can't. In this paper examination of various data mining strategies given in tables which were used and steady for clinical analysts or experts for exact heart disease finding.[[2](#_ENREF_2)]

# Related Work

Jaymin Patel[[3](#_ENREF_3)] mentioned, the biggest disease problem that causes death in the world is heart disease. Many researchers used different techniques to reduce the required test data using tree model algorithms and in this paper data mining techniques (random forest and tree algorithm) are used to extract hidden data. After analyzing the experimental results, the accurate and efficient model is J48 which has the highest accuracy that based on UCI and has less build model time.

Amin ul Haq mentioned[[4](#_ENREF_4)], it proves that the hybrid i

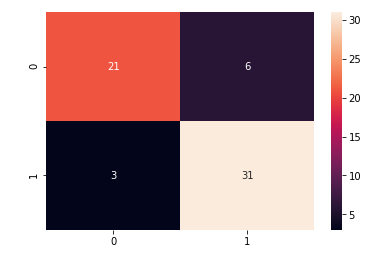
****

Figure 5: heat map representation of confusion matrices

Figure 7: heat map representation of confusion matrices

ntelligence machine learning-based system predicted heart diseased patient seven well-known classifiers are used with three feature selection classifier algorithms. The classifier logistic regression with 10% cross-validation has 89% accuracy which is best. Due to good performance, this system is the better predictive system in terms of accuracy.

Ayantan Dandapath mentioned [[5](#_ENREF_5)] ,In recent times scientists use machine learning methods headed for help health care and reliable physical and accurate data sets for diagnosis of disease. Various models used such as support vector machine, K nearest neighbor, random forest, and decision trees are very popular in researchers for diagnosis of disease. After analyzing the experimental results, the accurate and efficient model naïve Bayes classifier were fast and performed.

Beant Kaur [[6](#_ENREF_6)] mentioned By using the data mining technique, it takes less time for the prediction of disease with more accuracy. The result from using a nearly neural network is nearly accurate analysis shows that different technologies are used in different papers. The purpose is to make automated system to make heart prediction accurately.

Sonam Nikhar[[7](#_ENREF_7)] Researchers use many classifiers to predicting the disease; they make available a comprehensive explanation of naïve Bayes and decision tree classifier for more accuracy and efficiency. Decision tree and naïve Bayes proves the best classifier for this problem, after comparing both models the result shows that the decision tree has more accurate and efficient than naïve Bayes.

Abhishek Taneja [[8](#_ENREF_8)]As time passes the diseases also increase and the major disease that causes deaths due to heart disease so for predicting this disease the researchers use many algorithms such as random forest, naïve Bayes, J48 classifier for better accuracy and efficiency. After analyzing the experimental results, the accurate and efficient model J48 classifier was fast and performed well accurately, and gives 95.56% accuracy in predicting heart disease.

Youness Khourdif[[9](#_ENREF_9)] in this paper shows that the forecast of heart illness with machine learning techniques and optimization procedures has the advantage of dealing with complex nonlinear problems in this method FCBS method to filter and improve the quality of heart disease classification. Different experiments performed using different techniques and compare it with PSO and ACO approaches.

# Heart diseases

The heart is a vital organ of a human body part. It is just a siphon, which siphons blood through the body. Look if the flow of blood in the body is incompetent the parts of the body like the brain bear but if the heart stops working all around, death occurs in a matter of seconds. Life is thoroughly dependent upon the powerful working of the heart. The term Heart disease suggests a sickness of the heart and vein structure inside it[[10](#_ENREF_10)].

Various elements have been indicated that builds the danger of Heart sickness

* Heredity history
* Addicted to any drug
* Un prevailing diet
* Elevated blood pressure
* increased blood cholesterol
* body Obesity
* no physical exercise
* Hypertension

Elements like these are utilized to investigate Heart illness. By and large, the finding is for the most part dependent on the patient's present test outcomes and specialist's experience. Subsequently, a determination is an unpredictable undertaking that requires a lot of involvement and high expertise.[[10](#_ENREF_10)]

# Proposed Model

Today, numerous emergency clinics oversee medical care information utilizing medical care data frameworks; as the framework contains a tremendous measure of information, used to extricate shrouded data for making an astute clinical determination. The fundamental target our research is to engineer a Heart Disease Prediction system which will provide an analysis of heart illness using an realistic heart data set. To makeup this system, medical facts, for example, sex, pulse, and cholesterol like 13 figures ascribes are used. The information mining characterization strategies by using Logistic Regression, Naive Bayes, K-Nearest Neighbors, Decision tree, and Random Forest models[[4](#_ENREF_4)]

Following are subsections that briefly describe the working of our model

Figure 1 : steps to make data sets train and retriving results from the models after being evulated

1. **Dataset**

The dataset which we are using for the heart illness prediction includes information of 303 patients and with 14 features.

**5.1. Data Acquisition**

The dataset is collected from the Kaggle website

**5.2. Data Pre-Processing**

In this, we will be applying techniques to render out the abnormal data which may affect the accuracy rate afterwards. Coronary illness data is pre-processed after the collection of different records. The dataset covers an aggregate of 303 patient histories, where 6 records are for certain missing qualities. Those 6 records have been taken out from the dataset and the leftover 297 patient records are utilized in pre-preparing. The multi-class variable and double characterization are presented for the credits of the given dataset. The multi-class variable is utilized to pattern the presence or nonappearance of heart illness. In the case of the patient having a coronary illness, the worth is set to 1, else the worth is set to 0 signifying the shortfall of heart illness in the patient. The pre-preparing of information is done by changing over clinical records into determination esteems. The consequences of information pre-handling for 297 patient records demonstrate that 137 records show the estimation of 1 setting up the presence of heart illness while the excess 160 mirrored the estimation of 0 showing the shortfall of coronary illness.[[11](#_ENREF_11)]

**5.3. Attributes of Dataset**

Following are the main details of the features that are used in this paper.

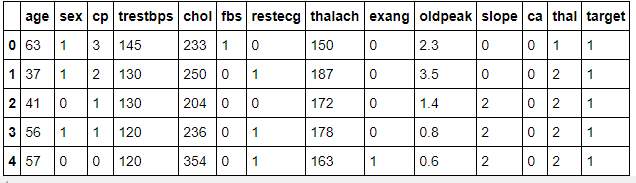


Figure 2: labels pf dataset used

* 1. **Dataset Attribute Information**

|  |  |
| --- | --- |
| **Features** | **Details** |
| Age | age |
| sex | 1: male, 0: female |
| Cp | chest pain type, 1: typical angina, 2: atypical angina,3: non-anginal pain, 4: asymptomatic |
| trestbps | resting blood pressure |
| chol | serum cholestoral in mg/dl |
| Fbs | fasting blood sugar > 120 mg/dl |
| restecg | resting electrocardiographic results (values 0,1,2) |
| Thalach | maximum heart rate achieved |
| exang | exercise-induced angina |
| Oldpeak | oldpeak = ST depression induced by exercise relative to rest |
| Slope | the slope of the peak exercise ST segment |
| Ca | number of major vessels (0-3) colored by flourosopy |
| thal | thal: 3 = normal; 6 = fixed defect; 7 = reversable defect |

Table 1: provides detailed information regarding the attributes used in the dataset

# Machine learning techniques used for prediction

**6.1. Logistic Regression**

Logistic regression is the classification technique, which helps to forecast the target variable’s probability. For example predicting age tells male or female, 0 or 1 so in simple words, dependent variables must have binary nature 0 or one that’s used for yes/no or failure/no, etc. Binary/binomial logistic regression is the simplest form of logistic regression that target/dependent variables have only 2 types of possibilities 0 or 1. and another type is multinomial logistic regression in which the dependent variables can have more than 3 possible unordered types. This paper uses different models, one is logistic regression.[[12](#_ENREF_12), [13](#_ENREF_13)]

The accuracy score achieved using Logistic Regression is: 85.25 %

**6.1.1. CONFUSION MATRICE**

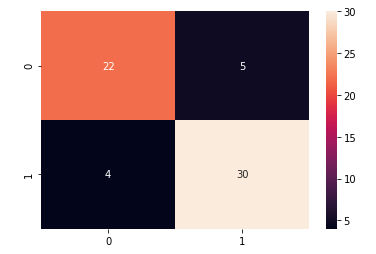


Figure 3: heat map representation of confusion matrices

Figure 4: the calculated value of confusion matrices is represented

**6.2. Naive Bayes**

Machine learning has a significant part in the medical services industry to empower health systems to appropriately utilize the data and analytics to distinguish feebleness that improves care with decrease costs. Naïve Bayes is a machine learning technique that utilizes the group of data that is supervised is used to precisely anticipate the aimed class for every instance of the provided figures. Heart illness sorting includes recognizing fit and sick people. Naive Bayes is generally constant regarding minor disparity or fluctuations in the datasets provide for training. [[14](#_ENREF_14), [15](#_ENREF_15)]

The accuracy score achieved using Naive Bayes is: 85.25 %

**6.2.1 CONFUSION MATRICE**

Figure 6: the calculated value of confusion matrices is represented

**6.3. K-Nearest Neighbors**

Supervised algorithms are utilized for the early prediction of heart illness. Nearest Neighbor (KNN) is the generally utilized lazy classification algorithm. KNN is a well-known, successful, and productive algorithm utilized to recognize a pattern. Clinical informational collections contain countless features. So, The Performance of the classifier will be decreased if the informational indexes contain noisy features.

The accuracy score achieved using KNN is: 67.21 %

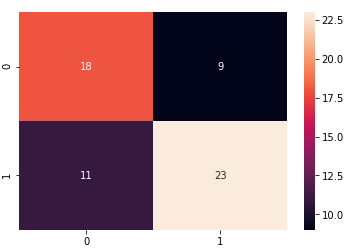
**6.3.1** **CONFUSION MATRICE**

Figure 8: the calculated value of confusion matrices is represented

**6.4. Random Forest**

Machine Learning algorithms and techniques are day by day enhancing and are expanding in the domain of health study by forecasting different illnesses. The use and concern of using machine learning prediction so we can adopt the thought process of an individual are building the new way which significant and flexible to forecast an illness. Random forest test of expanding the precision of Heart illness expectation has been elevated. Using this method of the random forest for our well-set attributes gave us more accuracy than the others. [[15](#_ENREF_15)]

The accuracy score achieved using Random Forest is: 95.08 %

**6.4.1 CONFUSION MATRICE**

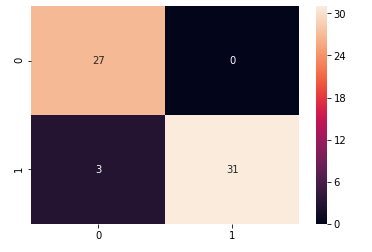


Figure 9: heat map representation of confusion matrices

Figure 10: the calculated value of confusion matrices is represented

**6.5. Decision trees**

A Decision Tree is perhaps the most famous and amazing characterization calculations in ML, which is for the most part utilized for foreseeing clear-cut data. Entropy/Information Gain and Gini Impurity are 2 key measurements utilized in deciding the pertinence of dynamic while building a choice tree model. The choice tree approach is all the more remarkable for arrangement issues. There are two stages in this procedure constructing a tree and applying the tree to the dataset. Pruning is a strategy that diminishes the size of the tree by eliminating overfitting information, which prompts helpless precision in forecasts. This method gives the most extreme exactness to preparing information. The general idea is to construct a tree that gives equilibrium of adaptability and exactness.

**6.5.1CONFUSION MATRICE**

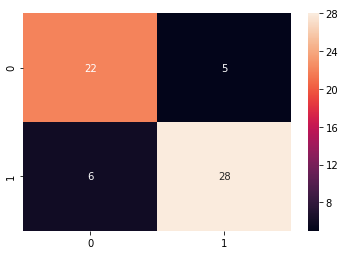


Figure 11: heat map representation of confusion matrices

Figure 12: the calculated value of confusion matrices is represented

# Results

|  |  |
| --- | --- |
| **Technique** | **Accuracy** |
| Logistic Regression | 85.25 % |
| Naive Bayes | 85.25 % |
| K-Nearest Neighbors | 67.21 % |
| Random Forest | 95.08 % |
| Decision Tree | 81.7% |

This paper uses different techniques and algorithms for better accuracy and Random Forest proves more accurate and efficient and it has an accuracy of 95.08%.

# Comparison of Models

Figure13:comparison among the model used in the paper

# Issue and challenges

Heart condition assurance is seen as a basic yet unpredictable endeavor that ought to be finished effectively and gainfully. By doing the manual process automated will be no doubt very helpful but on the other side very challenging as accuracy matters one wrong decision and person is gone. medical decisions that are made at the present only rely on the doctor’s skills and experience but not the data which has been taken before and is in the database stored unchecked[[7](#_ENREF_7)]. Due to this many of the lives are at the risk and relied on the doctor's choice what he may take after judgments of the physical appearance of the patients. Machine learning can drive out a new way of deciding with accuracy which baseness of one doctor.

# Conclusion

In this paper, the issue of obliging and summing up various calculations upon data, using machine algorithms utilized in the field of the clinical forecast is examined. The attention is on utilizing extraordinary calculations and mixes of a few objective credits for keen and powerful respiratory failure expectations utilizing Machine learning.[[6](#_ENREF_6)] We used different techniques and algorithms for better accuracy and in the end, Random Forest proves the most accurate technique for forecasting heart disease, after comparing different algorithms results, it shows that Random Forest provides accurate and efficient results with 95.08%.

1. **Future Work**

As future work, there still a need of working on Random Forest, there have to perform many experiments on more datasets and using algorithms for better accuracy and efficiency.

**Reference:**

1. Khourdifi, Y. and M. Bahaj, *Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization.* International Journal of Intelligent Engineering & Systems, 2019. **12**(1): p. 242-252.

2. Kaur, B. and W. Singh, *Review on heart disease prediction system using data mining techniques.* International journal on recent and innovation trends in computing and communication, 2014. **2**(10): p. 3003-3008.

3. patel, j., *Heart Disease Prediction Using Machine.* 2015.

4. haq, a.u., *A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms.*

5. dandpath, a., *Heart disease prediction using machine learning techniques.*

6. kaur, b., *Review on Heart Disease Prediction System using Data Mining Techniques.*

7. nikhar, s., *Prediction of Heart Disease Using Machine.*

8. Taneja, A., *Heart Disease Prediction System.*

9. Khourdif, Y., *Heart Disease Prediction and Classification Using Machine Learning Algorithms.*

10. Dangare, C.S., *Improved Study of Heart Disease Prediction System.*

11. Thirumalai, C.

12. Abdullah, A., et al., *Fake News Classification Bimodal using Convolutional Neural Network and Long Short-Term Memory.*

13. Alam, T.M. and M.J. Awan, *Domain analysis of information extraction techniques.* International Journal of Multidisciplinary Sciences and Engineering, 2018. **9**: p. 1-9.

14. Ali, Y., et al., *Detection of schistosomiasis factors using association rule mining.* IEEE Access, 2019. **7**: p. 186108-186114.

15. Awan, M.J., et al., *Efficient Detection of Knee Anterior Cruciate Ligament from Magnetic Resonance Imaging Using Deep Learning Approach.* Diagnostics, 2021. **11**(1): p. 105.