IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

USE OF MACHINE LEARNING ALGORITHMS FOR PREDICTION OF HEART DISEASE

G.V. Gayathri [*1], Md. Yazaz Rehman[*2]

[*1]Andhra University, Computer Science & Engineering Department, Anil Neerukonda Institute of Technology & Sciences, Visakhapatnam, Andhra Pradesh, India.

[*2]Andhra University, Computer Science & Engineering Department, Anil Neerukonda Institute of Technology & Sciences, Visakhapatnam, Andhra Pradesh, India.

ABSTRACT

Heart diseases are considered one of the most familiar causes of death worldwide. Early identification and medication can save a lot of people. There are multiple types of heart diseases that make it difficult to identify the type of disease that a patient is suffering with. Such data, whenever anticipated well ahead of time, can give significant instincts to specialists who can then adjust their conclusion and manage per patient premise. A machine must be developed such that it can identify the type of heart disease and update the machine itself by taking the experiences of the patients. (ML) can bring a compelling answer for navigation and precise forecasts. The clinical business is showing gigantic advancement in utilizing AI strategies. We work on foreseeing conceivable heart diseases in individuals utilizing Machine Learning calculations. This paper aims to build a model using multiple machine learning classifiers such as Logistic Regression (LR), Naive Bayes, Support Vector Machine (SVM), K-Nearest Neighbor (KNN) & Decision Tree. The model is able to implement hybrid classification by gaining weak and strong classifiers trained and tested on a dataset that contains multiple attributes and the efficient algorithm is considered the best. The decision tree performed best among the other algorithms with better accuracy and performance.

KEY WORDS:

Machine Learning, Logistic Regression (LR), Naive Bayes, Support Vector Machine (SVM), K-Nearest Neighbor (KNN) , Decision Tree.

1. INTRODUCTION

As per the World Health Organization, every year 12 million demises happen across the world because of heart disease. Coronary illness is perhaps the greatest reason for grimness and mortality among the number of inhabitants on the planet. The expectation of cardiovascular sickness is viewed as one of the main subjects in the segment of the data study. The heap of heart disease is quickly expanding all around the world in a couple of years. Many explorers have been directed in an endeavour to pinpoint the most persuasive variables of coronary illness as well as precisely foresee the general risk. Coronary illness is even featured as a silent death which prompts the demise of the individual without clear side effects.

The early conclusion of coronary illness assumes a crucial part in going with choices on the way of life changes in high-risk patients and in becoming decreases the entanglements. Machine learning ends up being viable in helping with decisions and predictions from the enormous amount of information created by the medical services industry. This paper expects to foresee future Heart Disease by breaking down information about patients who characterizes regardless of whether they have coronary illness utilizing machine learning methodology. Machine learning methods can be an aid in such a manner. Despite the fact that coronary illness can happen in various forms, there is a typical arrangement of important factors that impact regardless of whether somebody will eventually be in danger of coronary illness. Building a model using machine learning techniques can be cumulative as the model will be trained and tested with enormous of data and it can improve itself by taking future inputs. While building the heart disease predicting model the previous literature review papers helped us to acknowledge the improvements that are needed from the previously created models. The algorithms computed are Support Vector Machine (SVM), Naive Bayes, Decision Tree and Logistic Regression classifiers. By gathering the information from different sources, grouping them and applying an extensive examination, we can say that this strategy can be well indeed adjusted to do the prediction of coronary illness.

2. LITERATURE SURVEY

In this section we will mainly discuss about the background work that is carried out in order to prove the performance of our proposed Method. Literature survey is the most important step in software development process. For any software or application development, this step plays a very crucial role by determining the several factors like time, money, effort, lines of code and company strength. Once all these several factors are satisfied, then we need to determine which operating system and language used for developing the application. Once the programmers start building the application, they will first observe what are the predefined inventions that are done on same concept and then they will try to design the task in some innovated manner.

MOTIVATION

Santhana et al.,[1] The trees are built under a given set of conditions that provides either True or False decisions. Based upon dependent variables on vertical or horizontal split conditions the algorithms like SVM, and KNN provides the results. While building a decision tree the branches and nodes of the tree are built based upon the decisions that are made at every step this help the user to attain more knowledge on the dataset attributes. Cleveland dataset was used in this model. By using some methods, the training and testing data were split into 70% & 30% respectively. Decision Tree resulted in an accuracy of 91%. And for classification, the Naive Bayes approach was used. As the algorithm is well known for handling complicated linear and dependent data as it is most suitable for heart disease as both are familiar in nature. And after successful implementation, this algorithm achieved an accuracy of 87%.

Puroshottam et al.,[5] used the Cleveland dataset for pre-processing and training the model. And a model was built by implementing a decision tree and hill-climbing algorithms. An open-sourced data mining tool named Knowledge Extraction based on Evolutionary Learning (KEEL) was used. This tool helps in filling the missing values found in the dataset. As the decision tree follows a top-down approach, as each actual node is found it is selected a hill-climbing algorithm. This node is implemented by a test at each level. Confidence is the parameters and their values. The minimum confidence value used in the model is 0.25. And the model is able to provide an accuracy of 86.7%.

Senthil et al.,[3] proposed a model to improve the precision of identifying the heart disease affected patients. Using machine learning algorithms like KNN, Logistic Regression (LR), and Support Vector Machine (SVM), and combining them using Neural Networks produced an improvement in prediction accuracy of 88.7%. They also implemented the Hybrid Random Forest Linear Model (HRFLM) for better results.

Abhay et al.,[4] implemented a heart attack prediction system using the Deep learning method. Using the Recurrent Neural System, the model was able to predict the heart related infections in the patients. Implementing data mining and deep learning the model was able to provide precise results. This paper can helpful for building a hybrid model using deep learning and data mining techniques and it can be further improved by training the model on a real-time dataset of the patients.

Sonam et al.,[5] described a detailed explanation of a prediction model that works on decision trees and Naive Bayesian classifiers. The analysis led to the implementation of a prescient data mining approach on the same dataset. By using this hybrid approach decision tree gained the highest accuracy than the Naive Bayes.

3. EXISTING METHODOLOGY

In the existing system there was no proper method to identify the heart disease prediction at the early stages and hence there are several limitations present in the primitive days.

LIMITATION OF EXISTING SYSTEM

1. More Time Delay in order to predict the heart disease patients.
2. All the existing methods try to identify the noise by using manual approach.
3. There is no machine learning approach which can easily classify the patients into two categories.
4. There is no binary classifier which can distinguish the heart disease and normal patients separately.
5. For prediction of heart disease patients one should have strong knowledge to classify the heart disease based on symptoms.

4. PROPOSED MODEL

In this stage we are going to explain the proposed system and its model by using some model diagram and now we can clearly identify the step by step procedure of our proposed system.

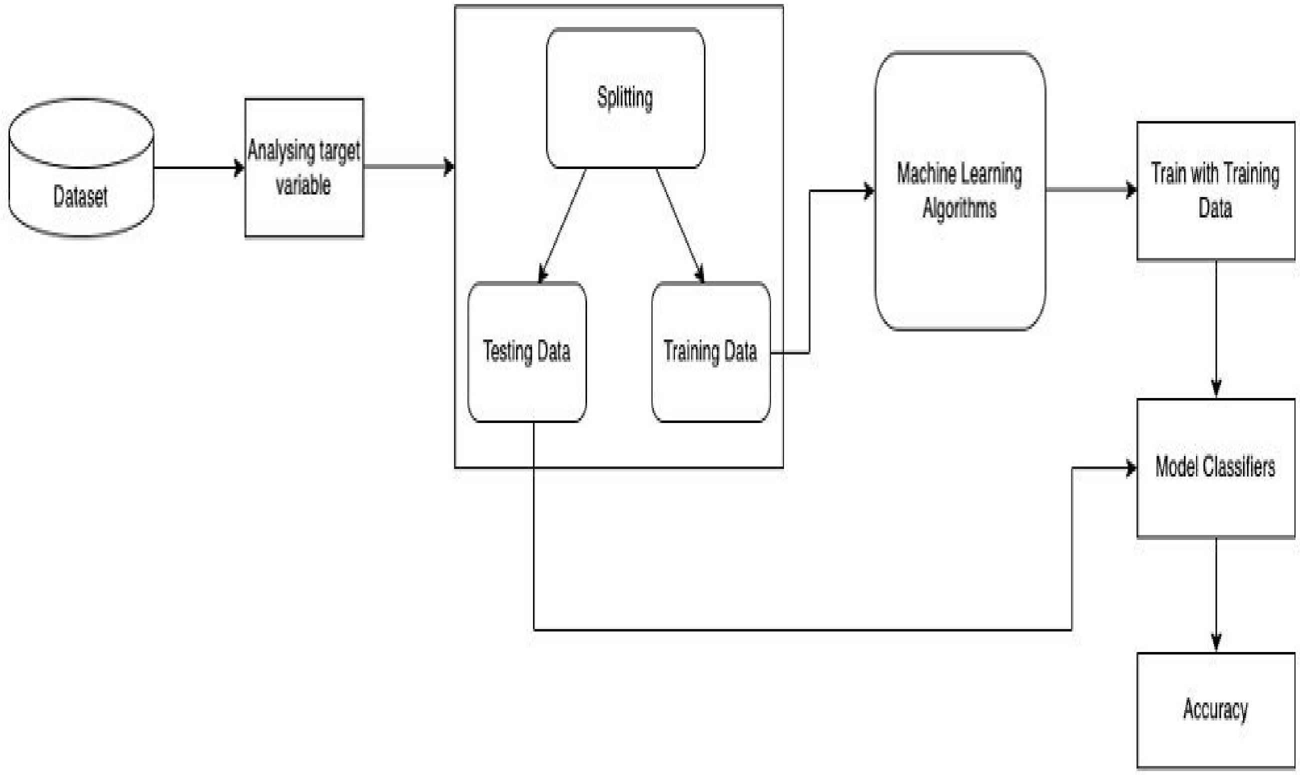


Figure 1: System architecture of the model

The proposed model is trained and tested on a vast amount of data from the dataset that is publicly available. The model is able to take input of age, sex, chest pain up to 4 types, resting blood pressure, serum cholesterol and it is able to predict whether the patient is infected with heart disease or not. This model is proposed to fulfill the following purposes:

1. The model is future proof such that new dataset changes can be trained and detected.
2. The model is trained and as well as tested further using various datasets.
3. The model is able to detect coronary illness.
4. To implement the system using Python Framework.
5. To obtain an efficient and accurate model using the algorithms mentioned.
6. To calculate and examine the accurate and time-efficient classifier among them.
7. To compare and contrast the results acquired from the dataset and select the best algorithm among them.

Sci-kit learns library is used for implementing the algorithms, this makes it convenient for data pre-processing, to be able to edit and produce more accurate results.

5. PROPOSED MACHINE LEARNING ALGORITHMS

The proposed system contains several machine learning algorithms and we try to compare all these algorithms in order to predict the heart disease of patients and then try to classify which one gives accurate result in order to predict the heart disease.

1) NAIVE BAYES CLASSIFIER:

The Naïve Bayes Classifier strategy is especially fit when the dimensionality of the data input is more. Regardless of its straightforwardness, Naive Bayes can frequently beat more modern classification strategies. The naive Bayes model recognizes the qualities of patients with coronary illness. It shows the likelihood of each input obtaining the predicted state.

2) LOGISTIC REGRESSION:

In Logistic regression, the target variable is taken, and it can be in a binary or discrete format i.e., either 1 or 0. It basically works on the sigmoid function so after successful calculation it results as 0 or 1, True or False etc. It works on mathematical functions, a complex function as a logistic or sigmoid function is calculated. This sigmoid function returns a value ranging between 0 and 1. And if the obtained result value is estimated to be less than 0.5 then it is considered as 0 and if it is greater than 0.5 and closer to 1 then it is considered as 1. Therefore, sigmoid plays a crucial role in building a model using logistic regression.

3) DECISION TREE:

A decision tree is a hierarchy constructed using information gain, Gini index, and leaf nodes, where the attributes with the highest information gain is considered. A tree is constructed by performing the probability of occurrence of each word in the email only after the pre-processing of the data is completed. The IG of the words is taken as a basic consideration and the tree is constructed, the leaf node represents the end of the decision tree.

The tree consists of the following nodes:

1. Decision Node: These are used for making decisions.
2. Leaf Node: These represent the outcomes of Decision nodes and do not have any branches.

4) SUPPORT VECTOR MACHINE(SVM):

A SVM model is an outline of the models as focused in space, planned with the goal that the instances of the discrete classes are partitioned. The points are isolated by a plane which is known as a hyperplane. A bunch of training data is given to it to check them as having a place with any of two classifications; an SVM training model then, at that point, constructs a model that relegates new instances of a similar space are planned and afterwards predicts to which classification they have a place, making it a non-probabilistic binary classifier.

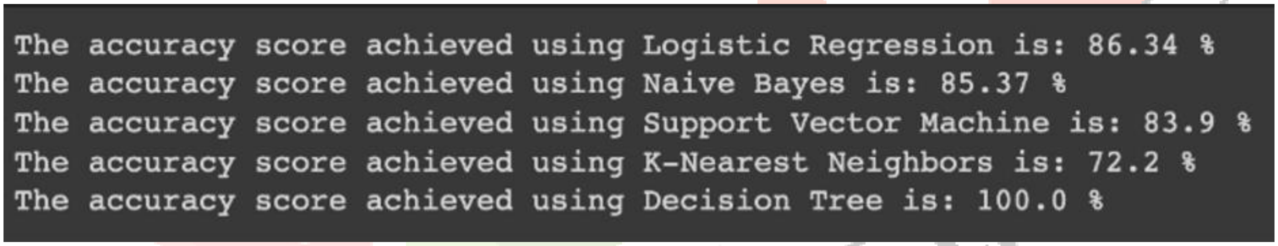
5) K-NEAREST NEIGHBOR (KNN):

The KNN algorithm is a simplified algorithm. The working of KNN is as follows:

1. Firstly, selecting the value for k.
2. Now the distance measure needs to be done. Using Euclidean distance, we can find the Euclidean distance neighbours for K.
3. By checking all the neighbors to this updated new point and see for the nearest neighbor to our desired point.
4. The class that is the highest the number is taken, and the maximum number is considered, and our new point is assigned to that class
5. By following the above steps, the KNN algorithm is implemented.

6. RESULT AND DISCUSSION

From the below two figures it can be seen that every machine learning classifier performed well on the dataset inserted into the machine. But among the five algorithms Decision Tree provided an accuracy of 100% and K-Nearest Neighbor (KNN) accuracy was low but it is considered then the existing models built on the KNN algorithm.



```
The accuracy score achieved using Logistic Regression is: 86.34 %
The accuracy score achieved using Naive Bayes is: 85.37 %
The accuracy score achieved using Support Vector Machine is: 83.9 %
The accuracy score achieved using K-Nearest Neighbors is: 72.2 %
The accuracy score achieved using Decision Tree is: 100.0 %
```

Figure 2: Accuracy score achieved from every machine learning classifier.

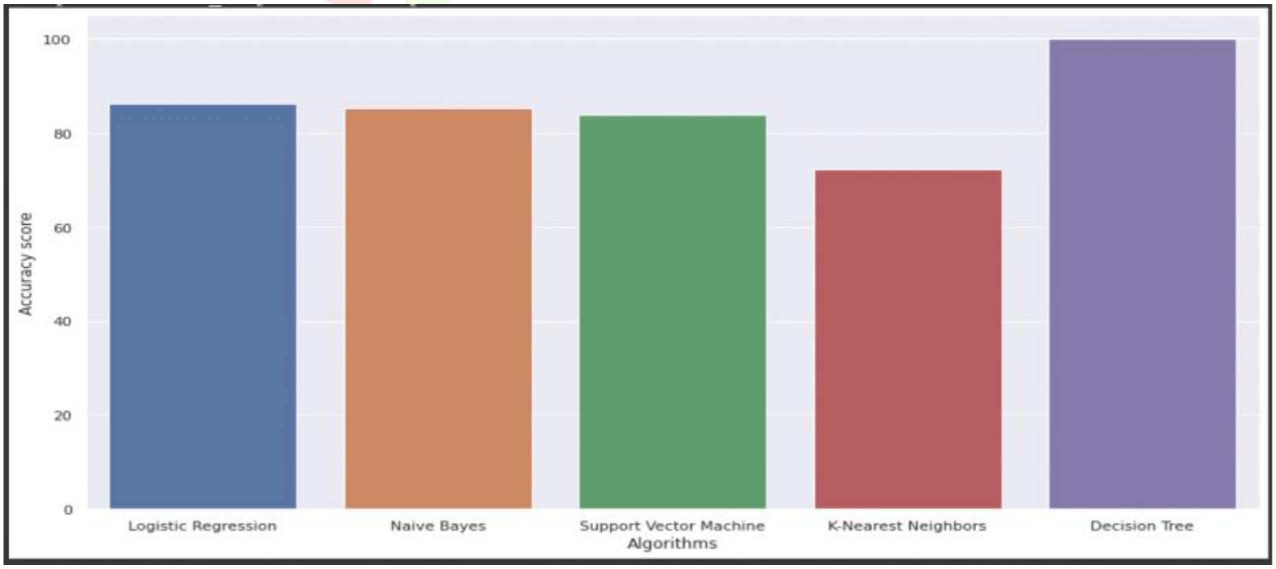


Figure 3: Graphical representation of accuracy from every machine learning classifier

7. CONCLUSION

The model is successfully built combined with machine learning algorithms. The dataset contains multiple attributes that were efficiently utilized to train and test this model. This paper provided accurate results distinguishing whether the patient is suffering from a coronary illness or not. All the algorithms were then tested with python libraries (sci-kit learn) and its modules. Thus, this paper provides accurate results with less time taken for computation than traditional heart disease prediction systems.

8. REFERENCES

[1] S. K. J. and G. S., "Prediction of Heart Disease Using Machine Learning Algorithms," 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT), 2019, pp. 1-5, doi: 10.1109/ICIICT1.2019.8741465.

[2] Abhay Kishore, Ajay Kumar, Karan Singh, Maninder Punia, Yogita Hambir., “Heart Attack Prediction Using Deep Learning,” International Research Journal of Engineering and Technology (IRJET), 2018, vol 5, no.4, April 2018.

[3]]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.

[4] Purushottam, K. Saxena and R. Sharma, "Efficient heart disease prediction system using decision tree," International Conference on Computing, Communication & Automation, 2015, pp. 72-77, doi: 10.1109/CCAA.2015.7148346.

[5] Nikhar, Sonam, and A. M. Karandikar. "Prediction of Heart Disease Using Machine Learning Algorithms." International Journal of Advanced Engineering, Management and Science, vol. 2, no. 6, Jun. 2016.