e-ISSN:2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:06/June-2021 Impact Factor- 5.354 www.irjmets.com

HEART ATTACK PREDICTION BY USING MACHINE LEARNING MODELS

Rohit Sharma*[1], Amit Kumar*[2], Tejbir Singh*[3]

*[1,2]B-Tech, Dept of CSE, GNIT, Mullana, Haryana, India.

*[3]Assist. Professor ,GNIT Mullana, Haryana, India.

ABSTRACT

Heart Attack is quite a common diseases nowadays. It is a cardiovascular disease that occurs when the flow of blood to the heart muscles is blocked. Medical science study has proven that the daily lifestyle we practice is the main reason behind this disease. There are also many key factors behind this disease. In this paper, we will use various machine learning algorithms like Logistic Regression, Naive Bayes, Random Forest Classifier, Extreme Gradient Boost, K-Nearest Neighbour, Decision Tree, Support Vector Machine on these key factors to analyze and predict the outcomes. Based on the accuracy provided by the various algorithm the most promising will be built on Extreme Gradient Boost as it gives the best Accuracy compared to other models. The main objective of this paper is to find the most promising model that can be trained and then use it to predict the outcomes with most accuracy.

Keywords: Heart Attack Prediction, Exploratory Data Analysis, Data Mining, Extreme Gradient Boost.

I. INTRODUCTION

Heart Attack is a very severe disease that can result in death. Nowadays, every year the number of people suffering from heart attack has significantly increased. Our present lifestyle is one of the reasons of it and people with problems like high cholesterol, blood pressure, diabetes etc., are more prone to fall for this disease. It has also been found that people with sophisticated diet or people with obesity are more likely to have a heart attack disease. Various symptoms can be observed but the most common are difficulty and shortness of breath, frequent pain in the chest muscles. The most common cause is the blood blockage in the heart muscles that results in heart attack. In this paper, we will try to analyse the key factor which cause a patient a heart attack. This can be achieved by training our model on data of past patients suffering from heart attack. We will use various machine learning algorithms like Logistic Regression, Naive Bayes, Random Forest Classifier, Extreme Gradient Boost, K-Nearest Neighbour, Decision Tree, Support Vector Machine on the dataset. Based on the accuracy provided by the various algorithm the most promising will be built on Extreme Gradient Boost as it gives the best Accuracy compared to other models.

II. LITERATURE SURVEY

[1]. Rajesh N, T Maneesha, Shaik Hafeez, and Hari Krishna have given a paper named “heart disease Prediction Using Machine Learning Models”. In this paper, they analyse the key factors causing heart diseases using different machine learning models

[2]. Costas Sideris, Nabil Alshurafa, Haik Kalantarian and Pourhomayoun have published a paper named, “Remote Health Monitoring Outcome Success prediction”. In this paper, they have portrayed an upgraded RHM framework, Wanda- CVD that is cell phone based and intended to give remote instructing and social help to members.

[3]. L.Sathish Kumar and A. Padmapriya has published a paper named “Prediction for similarities of disease using ID3 algorithm in television and mobile phone”. This paper helps us to understand a programmed and concealed approach to deal with the recognized designs that are covered up of coronary illness.

[4]. M.A.Nishara Banu and B.Gomathy has published a paper named, “Disease Predicting system using data mining techniques”. In this paper they portray about Maximal Frequent Item set algorithm and K-Means clustering. As classification is important for prediction of a disease.

[5]. Wiharto and Hari Kusnanto have published a paper named, “Intelligence System for Diagnosis Level of Coronary Heart Disease with K-Star Algorithm”. In this paper they portray an expectation framework for heart infection utilizing Learning vector Quantization neural system calculation.

III. MODELING AND ANALYSIS

All the models are compared based on accuracy below:

Table 1. Accuracy of models

	Model	Accuracy
0	Logistic Regression	85.245902
1	Naive Bayes	85.245902
2	Random Forest	86.885246
3	Extreme Gradient Boost	90.163934
4	K-Nearest Neighbour	88.524590
5	Decision Tree	81.967213
6	Support Vector Machine	88.524590

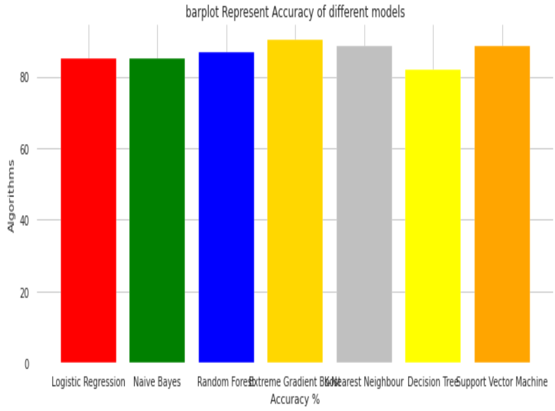

Fig.1 Accuracy Bar representation

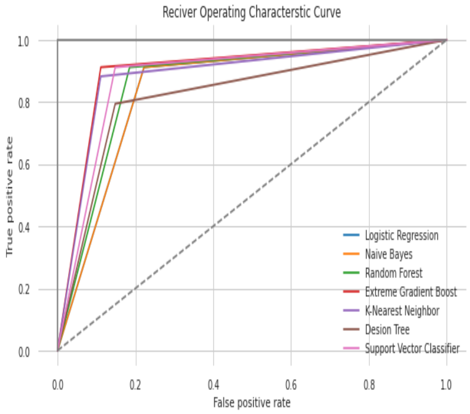

Fig.2 Receiver operating graph

IV. METHODOLOGY

Data Pre-Processing: The data available may have missing values which may lead to inconsistency and reduce the accuracy of the model. If there is any outlier they need to be removed and data need to be pre-processed. In this model we used fillna() to fill the missing values and we can also use dropna() to drop the missing value column so as to achieve consistency.

Logistic Regression: Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression [6] is estimating the parameters of a logistic model (a form of binary regression).

```
Accuracy of Logistic Regression: 85.2459
            0163934425
   precision   recall  f1-score  support
        0     0.88     0.78     0.82     27
        1     0.84     0.91     0.87     34
     accuracy                   0.85     61
   macro avg     0.86     0.84     0.85     61
```

```
weighted avg    0.85    0.85    0.85    6
                1
```

Naive Bayes: Naive Bayesian classifier frequently outflanks high order techniques which are complex. The Naïve Bayes treats every variable as independent which helps it to predict even if variables don't have proper relation [7].

```
Accuracy of Naive Bayes model: 85.24590
            163934425
     precision   recall  f1-score  support
        0    0.88    0.78    0.82    27
        1    0.84    0.91    0.87    34
     accuracy                0.85    61
   macro avg    0.86    0.84    0.85    61
weighted avg    0.85    0.85    0.85    61
```

Random Forest Classifier: The Random Forest Classifier is used on the train dataset and the following accuracy is obtained:

```
Accuracy of Random Forest: 86.88524590163934
     precision   recall  f1-score  support
        0    0.88    0.81    0.85    27
        1    0.86    0.91    0.89    34
     accuracy                0.87    61
   macro avg    0.87    0.86    0.87    61
weighted avg    0.87    0.87    0.87    61
```

Extreme Gradient Boost: The Extreme Gradient Boost Classifier is used on the train dataset and the following accuracy is obtained:

KNN: The KNN Classifier is used on the train dataset and the following accuracy is obtained:

Decision Tree Classifier: The Decision Tree Classifier is used on the dataset and the following accuracy is obta ined:

Support Vector Classifier: The Support Vector Classifier is used on the dataset and the following accuracy is o btained:

```
Accuracy of Support Vector Classifier: 88.5245901
            6393442
     precision   recall  f1-score  support
        0    0.88    0.85    0.87    27
        1    0.89    0.91    0.90    34
     accuracy                0.89    61
   macro avg    0.89    0.88    0.88    61
weighted avg    0.89    0.89    0.88    61
```

```
Accuracy of K-Neighbors Classifier: 88.52459
016393442
          precision   recall  f1-score   support
        0     0.86     0.89     0.87       27
        1     0.91     0.88     0.90       34
   accuracy                     0.89       61
  macro avg       0.88     0.89     0.88       61
weighted avg      0.89     0.89     0.89       61
```

```
Accuracy of DecisionTreeClassifier: 81.96721
              31147541
          precision   recall  f1-score   support
            0     0.77     0.85     0.81       27
            1     0.87     0.79     0.83       34
       accuracy                     0.82       61
      macro avg     0.82     0.82     0.82       61
   weighted avg     0.82     0.82     0.82       61
```

```
Accuracy of Extreme Gradient Boost: 90.1639
              344262295
          precision   recall  f1-score   support
            0     0.89     0.89     0.89       27
            1     0.91     0.91     0.91       34
       accuracy                     0.90       61
      macro avg     0.90     0.90     0.90       61
   weighted avg     0.90     0.90     0.90       61
```

V. RESULTS AND DISCUSSION

In order to increase the accuracy of the model we use ensembling. Here we use stacking technique on Extreme Gradient Boost, KNN and SVC. The accuracy obtained by the model is given below:

```
Accuracy of StackingCVClassifier: 91.8032
              7868852459
          precision   recall  f1-score   support
            0     0.92     0.89     0.91       27
            1     0.91     0.94     0.93       34
       accuracy                     0.92       61
      macro avg     0.92     0.92     0.92       61
   weighted avg     0.92     0.92     0.92       61
```

Results: The results of the model in predicting the Heart Attack disease gave us a wonderful accuracy 91.8%

VI. CONCLUSION

The analytical process started with data collection followed by data pre-processing where we cleaned the data by filling missing values, followed by exploratory analysis where we split the dataset into train and test sets and scaling was also done. After that data model which was created on Logistic Regression, Naive Bayes, Random Forest Classifier, Extreme Gradient Boost, K-Nearest Neighbour, Decision Tree, Support Vector Machine is applied on the training set and based on the test result accuracy and finally the model was built and accuracy was evaluated. The best accuracy on public test set is 90% by Extreme Gradient Boost. This helps us to understand following insights about the disease. Exercise induced angina, Chest pain is major symptoms of heart attack and play a major role in prediction. Ensembling technique can increase the accuracy of the model to 91%.

VII. REFERENCES

[1] Rajesh N, T Maneesha, Shaik Hafeez, Hari Krishna, "Heart disease Prediction Using Machine Learning Models" in International Journal of Engineering & Technology

[2] Nabil Alshurafa, Costas Sideris, Mohammad Pourhomayoun, Haik Kalantarian, Majid Sarrafzadeh "Remote Health Monitoring Outcome Success Prediction using Baseline and First Month Intervention Data" in IEEE Journal of Biomedical and Health Informatics

[3] Ponrathi Athilingam, Bradlee Jenkins, Marcia Johansson, Miguel Labrador "A Mobile Health Intervention to Improve Self-Care in Patients With Heart Failure: Pilot Randomized Control Trial" in JMIR Cardio 2017, vol. 1, issue 2, pg no:1

[4] DhafarHamed, Jwan K. Alwan, Mohamed Ibrahim, Mohammad B. Naeem "The Utilisation of Machine Learning Approaches for Medical Data Classification" in Annual Conference on New Trends in Information & Communications Technology Applications - march2017

[5] Applying k-Nearest Neighbour in Diagnosing Heart Disease Patients Mai Shouman, Tim Turner, and Rob Stocker International Journal of Information and Education Technology, Vol. 2, No. 3, June 2012

[6] "Logistic Regression Relating JAMA Patient Characteristics to Outcomes". Tolles, Juliana; Meurer, doi:10.1001/jama.2016.7653. ISSN 0098- 7484.

[7] Sonam Nikhar, A.M. Karandikar "Prediction of Heart Disease Using Machine Learning Algorithms" in International Journal of Advanced Engineering, Management and Science (IJAEMS) June2016 vol-2.