Self-Analysis of Heart Disease using Machine Learning

K. Reethu Priya[1], P. Ramya[2], S.R. Pavithra[3], Prof. M. Mohamed Yaseen[4]

[1,2,3]UG Student, Department of ECE, KCG College of Technology, Chennai-97
[4]Assistant Professor, Department of ECE, KCG College of Technology, Chennai-97

Abstract - *The purpose of this project is to foresee the heart disease so as to escape the aftermath. As diagnosis of this disease takes long hours manually, the automated diagnostic system for early heart disease detection needs to be developed. Data mining techniques contribute significantly to the production of such system. We have used classification techniques of machine learning in which the machine is learned from the past data and can predict the category of new input. This paper is a relative study on model implementation using K-Nearest Neighbor (KNN) algorithm. The efficiency of algorithm is calculated and compared with the results of accuracy, precision, sensitivity, specificity and False Positive Rate. The coding is done in python and executed in jupyter notebook.*

Key Words: **Algorithm, Machine Learning, K-Nearest Neighbor, Diagnosis, Classification**

1. INTRODUCTION

The prevalence of heart disease and stroke has increased over by 50% in the past 25 years. According to the World Health Organization (WHO) report, 17.5 million total world deaths result from heart attacks and strokes. Computer Science & Technology is used for diagnosing heart disease in bioinformatics and biomedicine. This can be further guided to a field called Machine Learning to predict heart disease based on standard data sets gathered. The datasets may have been recorded by few repositories in the world. Only we need to apply some classifiers of Machine Learning Techniques to signify the heart disease in a human. In this paper, we surveyed the research papers to compare the accuracy of various Machine Learning algorithms about heart disease based on the data sets given and their attributes. Hence, early detection of cardiac problems and methods for detecting heart disease can save a great deal of life and help doctors devise an effective treatment plan that eventually reduces the mortality rate due to cardiovascular diseases.

2. LITERATURE SURVEY

J Thomas, R Theresa Princy [1] made use of K nearest neighbour algorithm, Neural network, Naïve Bayes and Decision tree for heart disease prediction. This paper includes the survey on different classification methods used to predict each person's risk level based on age, gender, blood pressure, cholesterol, pulse rate. Risk level accuracy is high when a greater number of attributes are used.

Santhana Krishnan.J, Geetha S [2] made use of two supervised data mining algorithm was applied on the dataset to predict the possibilities of having heart disease of a patient, were analyzed with classification model namely Naïve Bayes Classifier and Decision tree classification. These two algorithms are applied to the same dataset in order to analyze the best algorithm in terms of accuracy. The Decision tree model has predicted the heart disease patient with an accuracy level of 91% and Naïve Bayes classifier has predicted heart disease patient with an accuracy level.

Senthilkumar Mohan, Chandrasegar Thirumalai, Gautam Srivastava [3] made use of machine learning techniques to process raw data and provide a new and novel discernment towards heart disease. They suggested hybrid HRFLM method to combine the Random Forest (RF) and Linear Method (LM) characteristics. HRFLM has proved to be predicting heart disease very accurately.

Rifki Wijaya, Ary Setijadi Prihatmanto, Kuspriyanto [4] discussed the development of heart disease prediction using Artificial Neural Network. They have taken 13 variables that can determine heart disease. Predicting a person's heart disease, a year ahead is done by analyzing heart rate data from the experiment. Using tools such as smart mirror, smart mouse, smart phones and a smart chair, data is collected. Heart rate data were stored and collected on a server through the Internet.

Umamaheswari K, Kiruba R [5] used various techniques involving the feature selection and classification of the heart disease resulting in accurate prediction. New algorithms and techniques involving ensemble methods involve multiple learning algorithms and hybrid systems that use the combination of Artificial Intelligence methods and techniques provides better accurate results.

Tulay KarayOlan, Ozkan Kilic [6] proposed a heart disease prediction system which uses artificial neural network backpropagation algorithm. They have taken 13 Clinical features as inputs to the neural network, and the neural network was then trained with a backpropagation algorithm to predict the absence or existence.

3. METHODOLOGY

We obtained the wide variety of real-life datasets from the Kaggle and used jupyter notebook as the platform for the purpose of coding. We have compared our performance against the K-Nearest Neighbor algorithm. The block diagram for heart disease prediction mode is shown below.

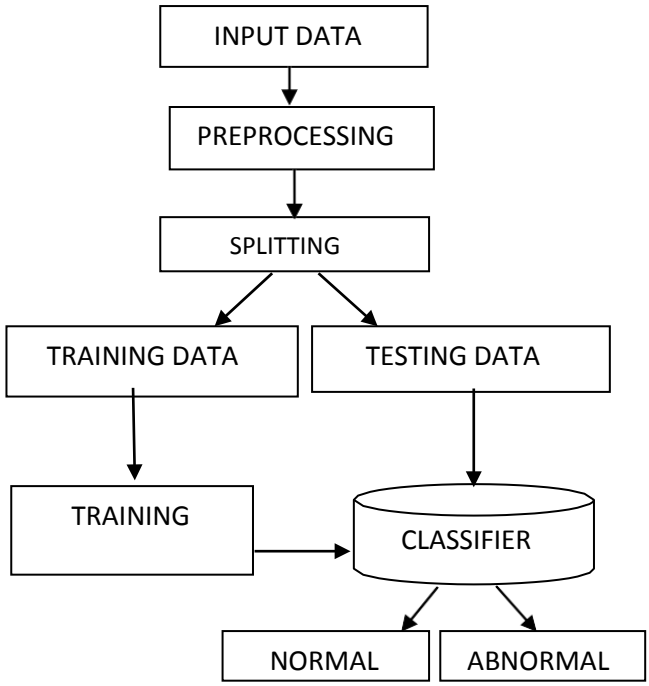

Fig -1: Heart Disease prediction model

3.1 Input Data

The data sets are collected from the kaggle and stored in the CSV file format. The machine learning model will be built using the data from the CSV file.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
5	57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
6	56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
7	44	1	1	120	263	0	1	173	0	0.0	2	0	3	1
8	52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
9	57	1	2	150	168	0	1	174	0	1.6	2	0	2	1
10	54	1	0	140	239	0	1	160	0	1.2	2	0	2	1
11	48	0	2	130	275	0	1	139	0	0.2	2	0	2	1
12	49	1	1	130	266	0	1	171	0	0.6	2	0	2	1
13	64	1	3	110	211	0	0	144	1	1.8	1	0	2	1
14	58	0	3	150	283	1	0	162	0	1.0	2	0	2	1

Fig -2: Dataset

3.2 Preprocessing

The dataset may contain NaN values. The NaN values cannot process by the programming hence these values need to convert into numerical values. In this approach mean of the column is calculated and NaN values are replaced by the mean. The aim of pre-processing is an improvement of the data that suppresses unwanted distortions or enhances some image features important for further processing.

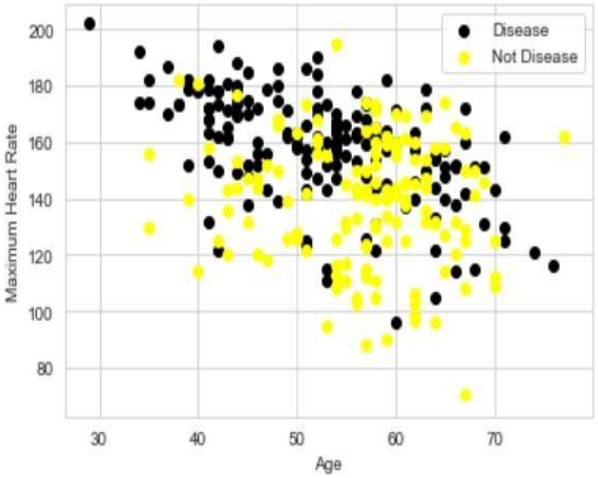

Fig -3: Classification of disease based on age

3.3 Splitting

The entire dataset is divided into a dataset for training and testing. For training the 80% data is taken while the remaining 20% data is used for testing.

Train_test_split

```
from sklearn.model_selection import train_test_split
y = df['target']
X = df.drop(['target'], axis = 1)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.33, random_state = 42)
```

Fig -4: Splitting of dataset

3.4 Model Selection

The most exciting step in designing any model of machine learning is algorithm selection. To large datasets we can use more than one kind - is the process by which the machine is trained on data which is well labeled for input and output. The model will learn about the training data and will be able to process future data to predict outcomes. They are grouped to Regression and Classification techniques. Unsupervised learning provides the machine with information that is neither classified nor labeled and allows the algorithm to analyze the information given without providing any directions. In unsupervised learning algorithm the machine

is trained from the data which is not labeled or classified making the algorithm to work without proper instructions. In our dataset we have an output variable or Dependent variable i.e. Y -having only two set of values, either 0 (Normal) or 1(Abnormal). So, Classification algorithm of supervised learning is applied on it.

3.5 K-Nearest Neighbor (KNN)

KNN is a non-parametric machine learning algorithm. The KNN algorithm is a supervised learning method. This means that all the data is labelled and the algorithm learns to predict the output from the input data. It works well even if the training data is large and contains noisy values. The data is divided into training sets and test sets. The train set is used for model building and training purposes. The k value is determined which is often the square root of the number of observations. Now, the test data is expected to be based on the model. For continuous variables, Euclidean distance, Manhattan distance and Minkowski distance measures may be used. Nevertheless, the measure widely used is the Euclidean distance.

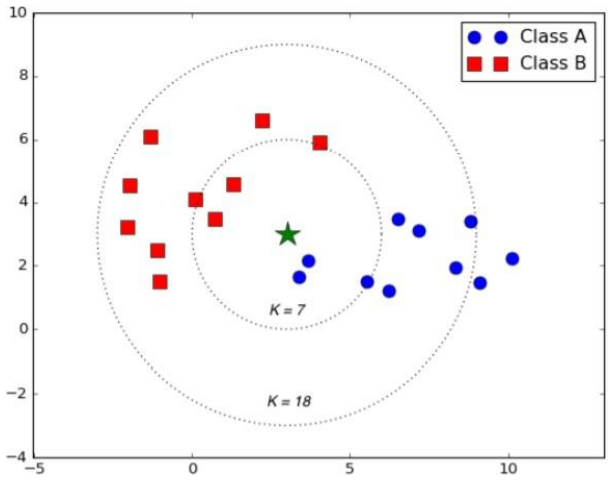

Fig -5: K Nearest Neighbor

Referring to Figure 5, Green star is the data point to be classified, blue circles are class A data points and red squares are class B rectangles. The Euclidean distance between the green star and all other red and blue points are measured. The star will be classified to the data points which have least distance. If k =7, then distance between all the seven points are measured from the star and star will be classified to the data points with least distance, in this case with blue data point.

4. PROPOSED MODEL

4.1 DATA DESCRIPTION

The dataset used here is taken from the kaggle for predicting heart disease. Kaggle has a collection of databases used to apply algorithms for machine learning. The dataset we are using here is a true dataset. The dataset consists of 300 data instances with the related 13 clinical parameters. The attributes taken for the model are shown in below table.

Table -1: Features used for the model

Attribute	Description
Age	Patient's age in years
Sex	Patient's gender (1=male; 0=female)
Cp	Chest pain type
Trestbps	Level of Blood Pressure at resting mode
Chol	Serum cholesterol at mg/dl
FBS	Fasting Blood Sugar >120mg/dl (1= true; 0= false)
Restecg	Resting electrocardiographic results
Thalach	Maximum heart rate achieved
Exang	Exercise induced angina (1= yes; 0= no)
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 fluoroscopy
Thal	Status of the heart (3-Normal; 6-Fixed Defect; 7-Reversablr defect)

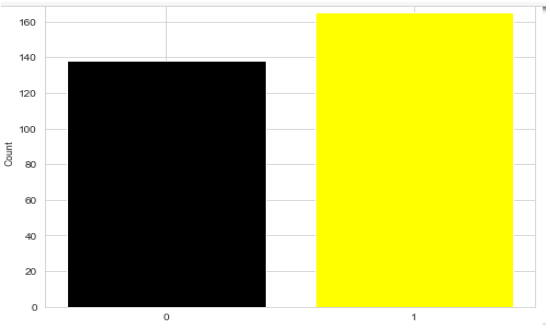

Fig -6: Plot showing class imbalance for machine learning, class inequality is a problem when the total number of positive and negative classes is not the same.

When the class inequality is not treated then it doesnot work well for the classifier. The above plot shows the unbalance for class.

4.2 RESULT AND DISCUSSION

Measuring accuracy is an important activity at machine learning. We have obtained the accuracy of 82% by using K-Nearest Neighbor algorithm. AUC-ROC curve is a performance measurement at different threshold settings for classification problem. ROC is a probability curve, and AUC is a degree or separability metric. The ROC curve is formed by the plotting at different threshold settings of the true positive rate (TPR) against the false positive rate (FPR). FPR and TPR define a ROC space as x and y axes, respectively, representing relative trade-offs between the true positive and the false positive. The TPR determines how many positive results are right among all positive samples available during the test. On the other hand, the FPR determines how many positive results are incorrect among all the negative samples available during the test. The diagonal divides the ROC space. Points above the diagonal are good results of classification and points below the line are bad results. It says how much model between classes can be separated.

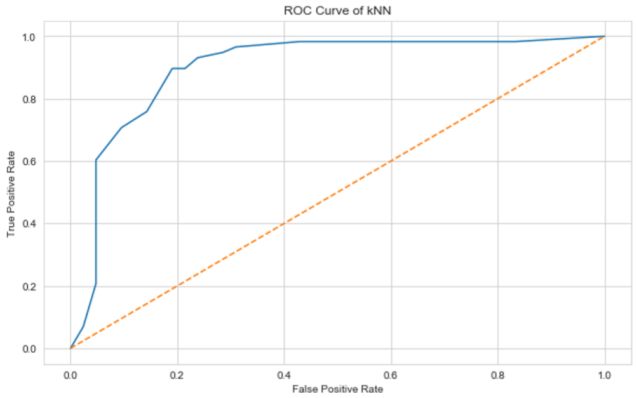

Fig -7: ROC Curve for KNN

5. CONCLUSION

In this paper, supervised data mining algorithm was applied on the dataset to predict the possibilities of having heart disease of a patient. The main objective of this paper is to provide an insight about prediction of risk of level of heart disease using machine learning algorithm with Grid Search for finding better accuracy. By predicting its occurrence earlier, the patient can be given treatment to prevent the heart disease. The developed system with the classification algorithm used for the machine learning can be used in the future to predict or diagnose certain diseases. The research may be extended or improved to simplify the study of heart disease including some other algorithms for machine learning.

REFERENCES

[1] J. Santhana Krishnan and S. Geetha, "Prediction of Heart Disease Using Machine Learning Algorithms," in International Conference on Innovation in Information and Communication Technology, 2019.M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[2] J. Thomas and R. Theresa Princy, "Human heart disease prediction system using data mining techniques," in International Conference on circuits,power and computing technologies, 2016.

[3] Senthilkumar Mohan, Chandrasegar Thirumalai and Gautam Srivastava, "Effective Heart Disease Prediction Using Hybrid," IEEE Access, vol. 7, 2019.

[4] Rifki Wijaya, Ary Setijadi Prihatmanto and Kuspriyanto, "Preliminary design of estimation heart disease by using," in Joint International Conference on Rural Information and Communication , 2014.

[5] S.P. Rajamhoana, C. Akalya Devi and K. Umamaheswari, "Analysis of Neural Networks Based Heart Disease," in International Conference on Human System Interaction(HSI), 2018.

[6] Tulay Karayilan and Ozkan Kilio, "Prediction of heart disease using neural network," in International Conference on Computer Science and Engineering, 2017.