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 DISEASE PREDICTION APPLICATION USING MACHINE LEARNING

Tarun Rahuja*1, Ms.Nidhi Sengar*2

*1B.Tech Scholar, Department of IT, Maharaja Agrasen Institute Of Technology, Delhi, India.

*2Assistant Professor, Department of IT, Maharaja Agrasen Institute Of Technology, Delhi, India.

ABSTRACT

With each passing day the instances of cardiovascular diseases have been increasing at a rapid rate. The primary reasons behind this situation are the increasing degree of stress and unhealthy lifestyles which includes lack of exercise, improper diet regimes and practices such as smoking. Owing to this, it is of paramount importance to come up with quick yet accurate predictive techniques that can be used to indicate to an individual if he or she is a possible victim of heart disease so that one can reach out to a licensed practitioner on time and commence the corrective measures to avert the danger of a cardiac arrest. To address this concern , a heart disease prediction system has been developed using predictive modeling techniques to estimate the likelihood of occurrence of heart disease in an individual given certain parameters as input. To model the system 2 datasets were used namely, Cleveland Dataset from UCI medical dataset repository and Framingham Dataset. Supervised learning algorithms like Naïve Bayes, K-Nearest Neighbours, decision trees and random forests were applied on the datasets to provide accuracies of up to 85%. To make the system interactive and accessible, a front end was developed for the end users to provide custom inputs to the trained model integrated with a python backend and get predictions. Such systems provide quick and economically viable solutions to health hazards like heart diseases.

Keywords: Supervised Learning, Predictive Modeling, Naïve Bayes, Random Forests, Decision Trees, K-Nearest Neighbours.

I. INTRODUCTION

"Machine Learning is a way of Manipulating and extraction of implicit, previously unknown/known and potential useful information about data" **[1]**. Machine learning is a vast area of scientific research that has seen tremendous growth in recent times. With the advent of machine learning in almost all spheres of life like finance, marketing, it is almost imperative to explore the potential of machine learning in healthcare industry. Machine learning encompasses various techniques of supervised, unsupervised and ensemble learning which can be used in classification use cases. This can be leveraged in our use case of Heart disease prediction to alert an individual about the plausibility of cardiac ailment so that one can seek medical attention before any significant damage occurs. Such systems are the need of the hour considering the steep rise in global cases of cardiovascular disease. World Health Organisation estimates that more than 10 million individuals succumb to cardiac ailments per annum **[2].** Multiple factors such as personal and professional routines and genetic makeup accredit for heart disease. Various habitual risk factors such as smoking, over consumption of alcohol, caffeine, stressful routines, and inadequate physical activity along with other health related conditions like obesity, hypertension, high blood cholesterol are the primary causes known for heart disease. Heart disease is one of the leading causes of death in adults and our system can be used to counter this, by asking the end users simple questions like whether they smoke, have had a history of hypertension/diabetes to alert them of a possible heart ailment. Timely awareness of the ailment followed by medical guidance is of paramount importance in treating diseases of the heart, which is where our system thrives to act. It provides transparency to the patient about the possibility of a heart diseases so that timely action can be taken up. The project makes use of multiple data mining techniques : (1) Logistic Regression (2) Support Vector Machines (3) Decision Trees (4) Random Forests (5) Naïve Bayes (6) K-Nearest Neighbours. To make it more user friendly 2 different datasets were used for training the model one (Cleveland Dataset) of which required input of advanced parameters such a fasting blood sugar, type of chest pain and the other one (Framingham Dataset) being relatively simple and requiring the entry of relatively simple data by the end user such if he or she smokes or not, if one has had a history of diabetes.

II. RELATED WORKS

There has been a significant amount of research that has been accomplished in the field of intelligent heart disease prediction using data mining techniques. AH Chen worked on a heart disease prediction system that can aid medical practioners in predicting the plausibility of heart disease based on the clinical data of patients. Thirteen important clinical features such as age, sex, chest pain type were selected. An artificial neural network algorithm was used for classifying heart disease based on these clinical features [3]. Mrudula Gudadhe presented a decision support system for heart disease classification. Support vector machine (SVM) and artificial neural network (ANN) were the two main methods used in this system[4]. Lamido Yahaya , Nathaniel David Oye, Etemi Joshua Garba in their paper, A Comprehensive Review on Heart Disease Prediction Using Data Mining and Machine Learning Techniques, analysed the state of various medical decision support systems for heart disease prediction, proposed by multiple researchers using data mining and machine learning techniques. Classification algorithms such as the Naïve Bayes (NB), Decision Tree (DT), and Artificial Neural Network (ANN) were used to predict heart diseases to obtain varying degrees of accuracy [5]. Asha Rajkumar worked on diagnosis of heart disease using classification based on supervised machine learning. Tanagra tool was used to classify the data. Tanagra is a data mining software used primarily for research purposes. It makes use of several data mining methods ranging from explanatory data analysis, statistical learning to machine learning and database domain [6].

III. METHODOLOGY

1) **Data Acquisition :** The data used as part of the project was collected from 2 different data sources :

2) **UCI Medical Repository :** The Cleveland Dataset that is a part of the UCI medical repository was used for the project. The dataset consists of 303 data points and 76 attributes out of which 14 attributes were chosen for the purpose of the research one of which was the target variable and the remaining 13 were predictor variables. The attributes.

3) **Framingham Dataset :** This dataset consists of 4241 data points each with 16 attributes. One of the attributes is the target variable while the rest are predictors.

4) **Exploratory Data Analysis :** The phase marks an in depth study of the datasets to figure out the most significant attributes that can be used as predictor variables. In the Cleveland dataset all the 13 predictor variables were moderately correlated to the target variable and the relative correlation between attributes was low, hence, all the variables were used as predictors whereas in the Framingham dataset, some attributes that were directly related to each other such as the Boolean variable to mark whether a person smokes or not and the number of cigerettes smoked per day are directly related so, only one can be picked. On similar lines , unrelated attributes such as education were dropped and the dataset was sanitized to contain only relevant predictor variables.

5) **Model fitting** : After the data was cleaned and analyzed to filter out the most significant attributes, the next step was to apply various supervised learning machine learning algorithms to the the dataset. The implementations used were from the sklearn library written in python. Algorithms used in the analysis were :

a) Logistic Regression: Logistic regression is a supervised learning classification algorithm used to assign labels to data instances. The class label is binary in nature, having data coded as either 1 (Class A) or 0 (class B). Though logistic regression can also be used for multi class classification, its most common use cases involve binary classification majorly. Logistic regression uses a sigmoid function which is as follows:

$$f(x) = \frac{1}{1+e^{-x}} \qquad - (1)$$

The input values are combined with weights similar to linear regression to produce an output but in logistic regression instead of using the predicted value, the sigmoid function takes in the value estimated by predictor variables and generates a probability. There is also a preset threshold value depending upon which, the value of the probability obtained from the sigmoid function is used to decide to which class the data instance belongs to.

b) K-Nearest Neighbours: KNN works by finding the distance between a new data point and all the classified data points in the input and chooses the specified number of labeled examples (K) closest to the unclassified instance, then picking up the most frequent label (in the case of classification use cases) or taking the statistical mean of the labels (in the case of regression problems). The distance chosen can differ based upon implementation. One may make use of minkowsky, Euclidean distance, cosine similarity and chi-square as distance measures of the new, yet to be classified point from its neighbours.

c) Naïve Bayes : Naïve Bayes algorithms is a supervised machine learning algorithm used in classification problems. It is based upon Conditional probability(Baye's theorem) approach with an assumption that all the input variables are independent of each other. This model is useful in cases wherein the number of data points available is sufficiently large but the number of attributes are less and the prediction is to be achieved in a short span of time.

Mathematically denoting,

$$P(\text{class}|\text{data}) = \frac{(P(\text{data}|\text{class}) * P(\text{class}))}{P(\text{data})} \quad - (2)$$

d) Support Vector Machines : In the SVM algorithm, we plot each point in the domain as a point in n-dimensional space, n being the number of columns/attributes in the dataset. Then, we perform classification by finding the hyperplane that differentiates data points into various classes. Hyperplanes are decision boundaries that help in differentiating the data points into various classes. The dimension of the hyperplane depends upon the number of attributes. In general for a d-dimensional feature space, the hyperplane is a function of degree (d-1) . Say, if the feature space has 3 attributes then the hyperplane is a plane.

e) Decision Trees :Decision trees work by progressively splitting the input data in different classes based upon the value of a particular attribute in the domain. The tree consists of two components, namely, leaves and decision nodes. The leaves are the final classes to which a data point belongs. The decision nodes are where the data is differentiated based upon the condition imposed on the value of a particular attribute. There are many algorithms used to construct Decision Trees, but one of the most used algorithm is the ID3 Algorithm (Iterative Dichotomiser 3).

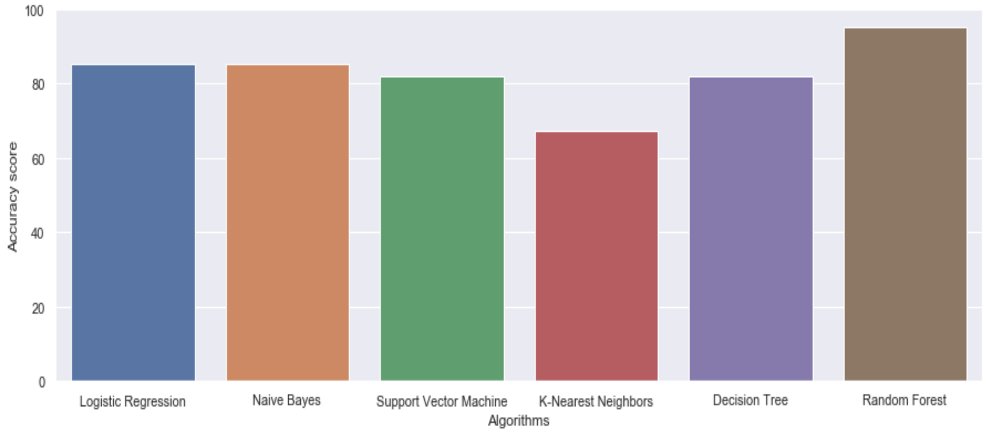
f) Random Forests : It is based on ensemble learning, which is a process of making use of multiple classifiers to make prediction, this solves the problem of overfitting. As the name suggests, in Random forests there are multiple decision trees operating in a parallel fashion upon different subsets of the dataset. It is advantageous over decision trees in terms of accuracy and efficiency since multiple decision trees act on the data in synchrony and the final classification is based out of the individual classifications of each decision tree.

4) Connecting to web layer : The trained machine learning models were interfaced with a web application whose backend was written using flask framework in python and the front end was designed using Vanilla JS. The front end consisted of 2 forms that allowed end users to select the model they wanted to interact with and depending upon their choice a form was displayed with fields that were required for the chosen model to function.

IV. RESULTS

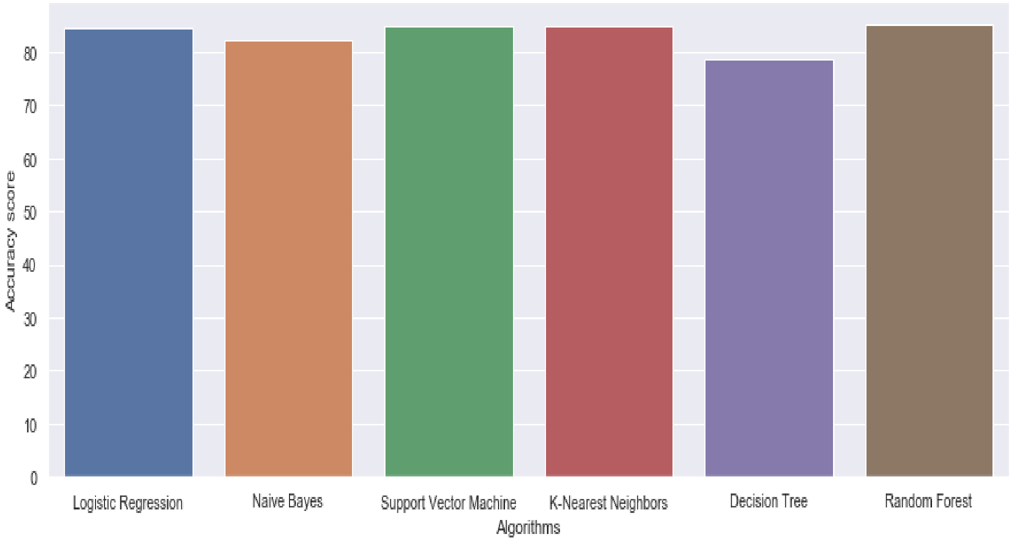
a) Cleveland Dataset

Algorithms Used	Accuracy achieved(in %)
Logistic Regression:	85.25
Naïve Bayes	85.25
K-Nearest Neighbours	67.21
Support Vector Machines	81.97
Decision Tree	81.97
Random Forest	95.08



b) Framingham Dataset

Algorithms Used	Accuracy achieved(in %)
Logistic Regression:	84.52
Naïve Bayes	82.24
K-Nearest Neighbours	85.06
Support Vector Machines	84.79
Decision Tree	78.60
Random Forest	85.15



V. CONCLUSION

This project has proved the viability of using machine learning and data mining techniques in the area of healthcare. The system developed has delivered accuracies upto 85 % with the Framingham dataset and 95 % with the Cleveland dataset. These accuracies are at par with the current standards of testing in place. Random forests algorithm performed well under the given configurations owing to the fact that it uses multiple decision trees in tandem and counteracts the problem of overfitting. Naïve Bayes and SVM also performed well in terms of accuracy and computational efficacy. This project can prove to be substantial in situations where an early diagnosis of heart disease can prevent mortality due to lack of awareness about illness. Such automation also cuts down upon human errors as part of pre-treatment. The integration of web layer with a clean and easy to interact user interface further enhances the accessibility and usability of this project.

VI. REFERENCES

[1] Soni J, Ansari U, Sharma D & Soni S (2011). Predictive data mining for medical diagnosis: an overview of heart disease prediction. International Journal of Computer Applications, 17(8), 43-8

[2] Dangare C S & Apte S S (2012). Improved study of heart disease prediction system using data mining classification techniques. International Journal of Computer Applications, 47(10), 44-8.

[3] AH Chen, SY Huang, PS Hong, CH Cheng, and EJ Lin,2011, "HDPS: Heart Disease Prediction System",Computing in Cardiology, ISSN: 0276-6574, pp.557- 560.

[4] Mrudula Gudadhe, Kapil Wankhade, and Snehlata Dongre, Sept 2010,"Decision Support System for Heart Disease Based on Support Vector Machine and Artificial Neural Network",International Conference on Computer and Communication Technology (ICCCT),DOI:10.1109/ICCCT.2010.5640377, 17-19.

[5] Lamido Yahaya, Nathaniel David Oye, Etemi Joshua Garba. A Comprehensive Review on Heart Disease Prediction Using Data Mining and Machine Learning Techniques. American Journal of Artificial Intelligence. Vol. 4, No. 1, 2020, pp. 20-29. doi: 10.11648/j.ajai.20200401.12

[6] Asha Rajkumar, and Mrs G. Sophia Reena, 2010, "Diagnosis of Heart Disease using Data Mining Algorithms",Global Journal of Computer Science and Technology,Vol. 10,Issue 10, pp.38-43, September.