International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 04 | Apr 2020 www.irjet.net p-ISSN: 2395-0072

Classification Techniques for Heart Disease Prediction

Raksha[1], Shreya P[1], Vineetha[1]

[1]*Dept. of ISE, The National Institute of Engineering, Mysuru*

Abstract - *Machine Learning is used across many ranges around the world. The healthcare industry is no exclusion. Machine Learning can play an essential role in predicting presence/absence of locomotors disorders, Heart diseases and more. Such information, if predicted well in advance, can provide important intuitions to doctors who can then adapt their diagnosis and dealing per patient basis. Here we will be predicting possible Heart Diseases in people using Machine Learning algorithms. The dataset has been taken from Kaggle. The various attributes related to cause of Heart Disease are viz: gender, age, chest pain, blood sugar, blood pressure etc. that can predict early symptoms heart disease. Utilization of suitable technology support in this regard can prove to be highly beneficial to the medical fraternity and patients. The use of these two algorithms classifies output data as no, low, average, high and very high. The algorithms used are Naïve Bayes and Random Forest algorithm. The dataset has been taken from Kaggle. The various attributes related to cause of Heart Disease are viz: gender, age, chest pain, blood sugar, blood pressure etc. that can predict early symptoms heart disease. Utilization of suitable technology support in this regard can prove to be highly beneficial to the medical fraternity and patients. The use of these two algorithms classifies output data as no, low, average, high and very high. The algorithms used are Naïve Bayes and Random Forest algorithm.*

Key Words: Heart Diseases, Naïve Bayes, Random Forest Algorithm, Kaggle, Health Care.

1. INTRODUCTION

Use of computing in the field of medicine can be seen from the early 1950s (Lai, 2013). However, the first applications of AI in medicine can only be seen during 1970s through expert systems such as INTERNIST-I, MYCIN, ONCOSIN (Hanson, 2001; Lai, 2013).The application of artificial intelligence in medicine was mostly limited in the US before 1980. An international conference was organized on 13-14 September 1985 in Italy. With the aim of creating an active research community, in 1986 "Society for Artificial Intelligence in Medicine" was established which biennially organizes international conferences. (Peek et al., 2015)One of the problems associated with using artificial intelligence in medicine is unavailability of data, which can be solved by the use of Electronic Medical Records (EMR). The concept of EMR was introduced by Larry Weed during the late 1960s. The US government started using EMR in the1970s with the Department of Veteran Affairs. (Atherton, 2011) "Knowledge Engineering" was the major research theme during 1980s, but after 1990 the research shifted to various topics. After 2000, "Machine Learning and Data Mining" has been the major research theme. (Peek et al., 2015)

1.1 CURRENT SCENARIO

In many countries health records are being digitized. The adoption of EMR is also increasing. According to a data brief by The Office of National Coordinator for Health Information Technology (ONC), 3 out of 4 private or not-for-profit hospitals adopted at least a Basic EHR system in the US (Charles et al., 2015). In many other countries, different EHR systems exist. The Stockholm EPR corpus is a great example of such systems which consists data from 512 clinical units with over 2 million patient records (Dalianis, 2016). India is thinking about setting up a National eHealth Authority (NeHA) during the Digital India program (Ghosh, 2015).

These type of electronic health documents provide a huge amount of data for intelligence data analysis. Many researches have been conducted on predicting various diseases like Liver Disease, Heart Disease, Diabetes etc.,

detecting tumours, leukaemia etc. using computer vision, assisting doctors in making efficient decisions, which have been further discussed in background.

1.2 AIMS AND OBJECTIVES

1.2.1 AIMS

The primary aim of this paper is to analyze the "Cleveland Heart Disease Dataset" and use Naïve Bayes and Random Forest algorithm for prediction and develop a prediction engine. The secondary aim is to develop a web application that allows users to predict heart disease utilizing the prediction engine.

1.2.2 OBJECTIVES

The objectives set to achieve the aims of the content are:

- Research on statistical models in machine learning
- Obtain heart disease datasets and filter the data – to perform case study using statistical models.
- Develop functionality to allow users to submit news URLs.

2. CLASSIFICATION SUPERVISED LEARNING

Applications in which training data comprises of input vectors along with corresponding output/target vectors are known as supervised learning problems (Bishop, 2006). The majority of practical machine learning makes use of supervised learning. Suppose input variables x and output variable y, then an algorithm to learn the mapping function from x to y is

$y = (x)$ …equation-1

Equation 1 Mapping Function The goal of supervised machine learning is to approximate the mapping function so that value of y given new input data x can be found.

Why this content comes under classification problem and why it is supervised learning?

Since the project involves identifying whether or not there is risk of heart diseases, it is obviously a classification problem. The "Cleveland Heart Disease Dataset" datasets used for this project already have the outcome hence supervised learning.

3. CLASSIFICATION ALGORITHMS

Included in this paper are two supervised machine learning algorithms used during the project.

3.1 NAÏVE BAYES

Naive Bayes is a set of supervised learning algorithms based on the Bayes' theorem with the "naïve" assumption of independence between every pair of features (scikit-learn developers, 2016a). Despite of its simplicity, it often outperforms more sophisticated classification methods.

If there are input variables x and output variable y, Bayes' theorem states the following relationship.

$P(y|x)=(P(y).P(x|y))/ P(x)$…equation-2

Equation 2 Bayes' Theorem In this project, Gaussian Naïve Bayes algorithm has been implemented. In case of Gaussian Naïve Bayes, the likelihood of the features is assumed to be Gaussian i.e. all continuous values x associated with class y are distributed according to Gaussian distribution.

3.2 RANDOM FOREST

Random Forest is a supervised learning algorithm. Like you can already see from its name, it creates a forest and makes it somehow random. The forest it builds, is an ensemble of Decision Trees, most of the time trained with the "bagging" method. The general idea of the bagging method is that a combination of learning models increases the overall result. One big advantage of random forest is, that it can be used for both classification and regression problems, which form the majority of current machine learning systems. We will talk about random forest in classification, since classification is sometimes considered the building block of machine learning.

4. SIMILAR SYSTEMS

4.1 KAGGLE

Kaggle (Kaggle.com) is a predictive modelling and analytics competitions platform where companies, students, researchers, statisticians compete to produce the best models. Kaggle launches different competitions through which companies recruit people for jobs. Kaggle also allows users to share datasets, kernels, and discussion in the forum, launch competitions and the post jobs. Some of the above features are free whereas others are paid.

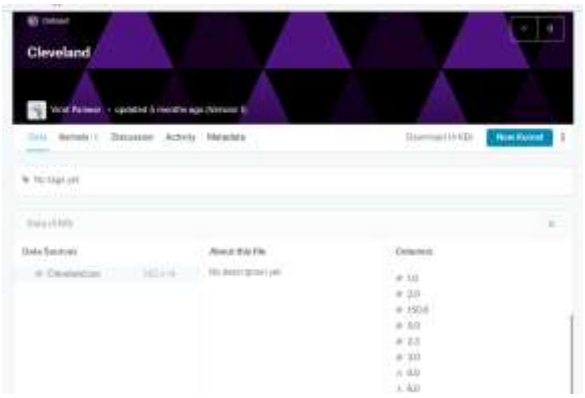

Fig 1- Cleveland Heart Disease Data set from Kaggle

4.2 UCI DATA REPOSITORY

(http://archive.ics.uci.edu/ml/) contains 351 datasets maintained by University of California, Irvine. It allows users to browse through datasets, download datasets and donate datasets.

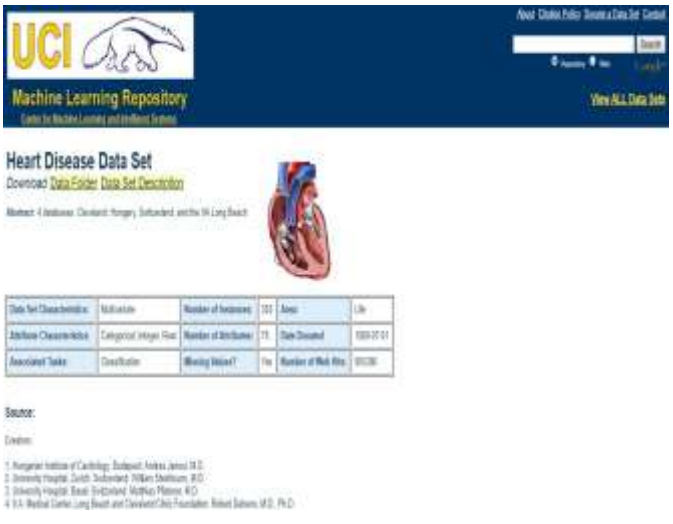

Fig 2- UCI Data Repository

5. IMPLEMENTATION-PREDICTION ENGINE

Here is the description of the datasets used in the development of prediction engine followed by the iterative implementation of prediction engine.

5.1 DATASETS

In this section, the "Cleveland Heart Disease" datasets have been discussed.

5.2 THE CLEVELAND HEART DISEASE DATASET

Name	Type	Description
Age	Continuous	age: age in years
Sex	Discrete	sex: sex (1 = male; 0 = female)
Cp	Discrete	chest pain location (1 = sub sternal; 0 = otherwise)
trestbps	Continuous	resting blood pressure (in mm Hg on admission to the hospital)
Chol	Continuous	serum cholestoral in mg/dl
fbs	Discrete	(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
Restecg	Discrete	resting electrocardiographic results (0,1,2)
Thalach	Continuous	maximum heart rate achieved
Exang	Continuous	exercise induced angina (1 = yes; 0 = no)
oldpeak	Discrete	ST depression
		induced by exercise relative to rest
Slope	Continuous	

Table 1 Dataset Attributes

The Cleveland Heart Disease Dataset has been taken from the UCI machine learning repository (Lichman, 2013). The dataset contains 303 records. The dataset contains 76 attributes but all the studies have only considered a subset of 14 attributes.

Fig 3 Receiver operating characteristics

Correlation Matrix is basically a covariance matrix. Also known as auto-covariance matrix, dispersion matrix, variance matrix, or variance-covariance matrix. It is a matrix in which i-j position defines the correlation between the i^{th} and j^{th} parameter of the given data-set.

6. IMPLEMENTATION (WEB APPLICATION)

This is regarding the implementation of web application following feature driven development. Each activity of feature driven development is discussed with artefacts produced during that activity.

6.1 HEART DISEASE PREDICTION FORM

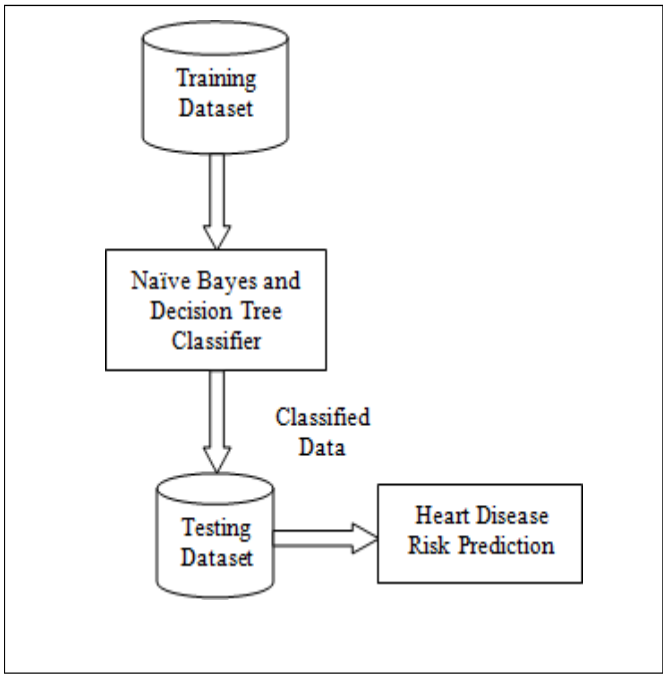

Fig 4 Sub flow diagram

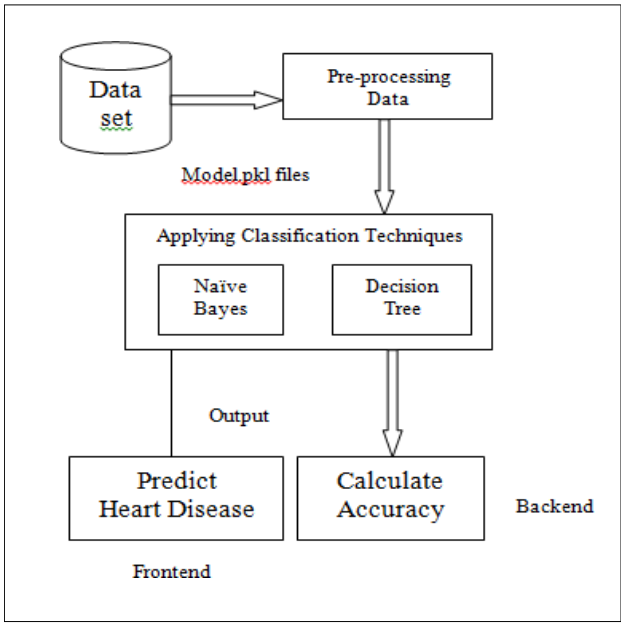

Fig 5 Main flow diagram

Ca	Continuous	number of major vessels (0-3) colored by fluoroscopy
Thal	Discrete	3 = normal; 6 = fixed defect; 7 = reversible defect
Num	Discrete	diagnosis of heart disease (angiographic disease status)

During this activity of feature driven development, wireframes were developed and software requirement specification document was prepared for capturing the requirements. ER Diagram was designed using the wireframes and requirement specification document. After that, for the completion of this activity, a domain object model was prepared along with the overall application architecture.

Sequence diagrams were then created as shown below in Figure 6

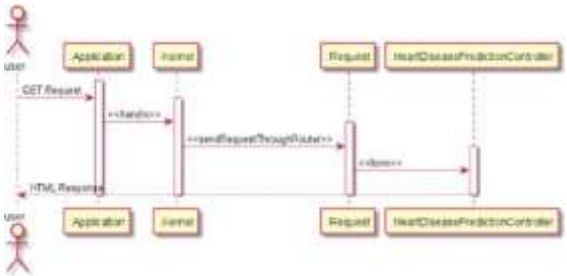

Fig 6 Sequence diagram- Heart disease prediction form

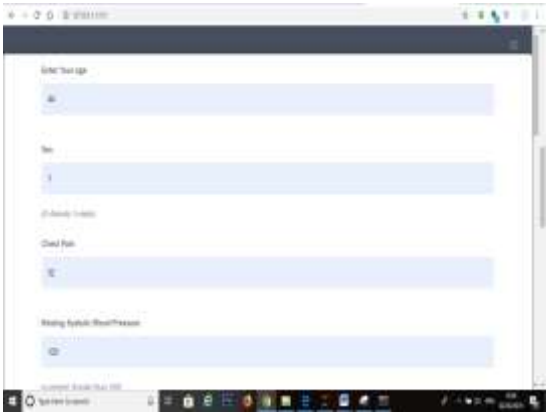

Fig 7 Prediction page navigator

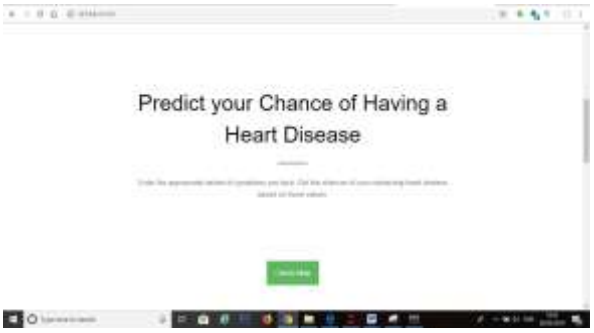

Fig 8 Prediction form

6. CONCLUSION

Here we presented a comparison of Multi-Layer Perceptron Neural Networks classifier with Naïve Bayes and Random Forest. Overall. Multi-Layer Perceptron outperformed every other classifier but at the cost of being computationally expensive. K-Nearest Neighbors performed as good as Multi-Layer Perceptron with far less computational requirement. The solution (web application) provided is a workable solution for the data problem. Currently there is a static prediction engine that serves prediction results for the diseases. There is a possibility of extending the system, to allow end-users to write their own prediction engine, execute it and publish it. Evaluating the project life cycle has indicated that the chosen methodology has been thoroughly followed. Evaluation of the project against functional and nonfunctional requirements and system testing indicates that all requirements have been fulfilled. In summary, the primary and secondary aims of the project have been achieved but there is still room for improvement and further enhancement.

7. REFERENCES

1. Intelligent Heart Disease Prediction System Using Data Mining Techniques-Sellappan Palaniappan, Rafiah Awang 978-1-4244-1968-5/08/ ©2008 IEEE.

2. Intelligent Heart Disease Prediction System Using Data Mining Techniques-Sellappan Palaniappan, Rafiah Awang 978-1-4244-1968-5/08/ ©2008 IEEE

3. https://www.researchgate.net/publication/319393368_Heart_Disease_Diagnosis_and_Prediction_Using_Machine_Learning_and_Data_Mining_Techniques_A_ReviewJ