Heart disease Prediction System

Hemalatha K N[1], Keerthana M[2], Meghana H R[3], Afreen Taj[4], Bhuvana M[5]

[1,2,3,4,5]Dept. of Computer Science and Engineering, Atria Institute of Technology

---***---

Abstract - *Millions of people today are suffering from various diseases that can prove fatal. Diseases like cancer, heart diseases, diabetes etc. cause a lot of health problems which may even lead to death. Identification of such diseases in early stages can help solve a lot of conditions related to them. Diagnosis of diseases at the right time will be of great help. Therefore to help the diagnosis process many data mining and machine learning techniques can be used. Data mining helps in finding out patterns in the data collected, by utilizing data mining we can assess numerous examples which will be use in future to settle on wise frameworks and choices. Data mining alludes to different strategies for distinguishing data or the reception of arrangements dependent on learning and information extraction of these information with the goal that they can be utilized in different territories, for decision making and prediction calculation of probability. Present health industry has a lot of information about patient's history. The huge amount of data can be scrutinized using data mining techniques, later machine learning algorithms can be used for the prediction process. Thus the main focus of the system is to make use data analytics to predict the presence of the disease and level of disease among patients.*

Key Words: - Machine learning, regression, decision tree, random forest, naïve bayes

1. INTRODUCTION

Data mining consolidates measurable investigation, AI and database technology. Data mining has been connected in a few regions of restorative administrations such has disclosure of connections among analysis information and put away clinical information. Cutting edge medicinal conclusion is a composite procedure which requires exact patient data, many long stretches of clinical experience and a decent information of the restorative writing .The medical industry has collected a lot of information ,which regrettably has not been used for finding any hidden relationships and patterns so that doctors cam make streamlined decisions[1]. Doctors always depend on their understanding and experience to make decisions. However, doctors having great knowledge in all sub-sectors are highly unavailable resource. The physicians won't be able to diagnose diseases accurately. Diagnosing a sickness appropriately at an earlier degree is a completely difficult undertaking because of the complicated interdependence of more than one factors. Solutions are usually made in a sanatorium based on intuition and enjoy of doctors, and no longer at the rich knowledge records which are hidden in the database. This technique results in undesirable biases, errors and needless health care fees, which impacts the pleasant of services furnished to sufferers. Machine learning can be used to determine the automated end of diagnostic guidelines from the past descriptions, efficiently treat an affected person, as well as experts and professionals will assist and make the diagnostic system more reliable. Intelligent decision making systems are characterized as intuitive PC frameworks to help settle on choices in the utilization of informational collections and models to discover issues, take care of issues and decide intelligent choice help structures are defined as interactive laptop structures to help make decisions in the use of records units and fashions to find issues, clear up issues and make selections. The proposed machine makes use of the analysis to combine and make the right choice on the health facility with a pc gadget. This affected person file can lessen the range of patients to improve the safety of scientific selections mistakes, reduces undesirable adjustments in practice and enhance affected person consequences. This thought is promising, as modeling and evaluation devices, such as information mining, have the capability to generate information-wealthy surroundings which could help to seriously enhance the caliber of medical choices.

2. SYSTEM DESIGN and IMPLEMENTATION

A. Architecture

The system makes use of machine learning algorithms to analyze the data available, train the models which are later evaluated. Algorithms used for prediction purpose are decision tree, naive bayes, random forest and linear regression. Models are built using all the four algorithms and their accuracies are compared. These models can be used to predict the type of heart disease patient is suffering from. System makes the prediction in the form of extended two level classification. Absence pf heart disease will be indicated by type 0 and presence will give values ranging from type 1 to type 4 [2] depending on the severeness of the disease. Algorithms make use of the Cleveland dataset to train the model. The dataset consist of twelve attributes and 801 instances. Attributes that are considered are: age, sex, chest pain, blood pressure, serum cholesterol, fasting blood sugar, electrocardiograph, max heart rate, ST_depression, slope and vessels [3]. Fig 1 represents the block diagram of the system. The dataset that is used is collected from Cleveland Clinic Foundation for heart disease.

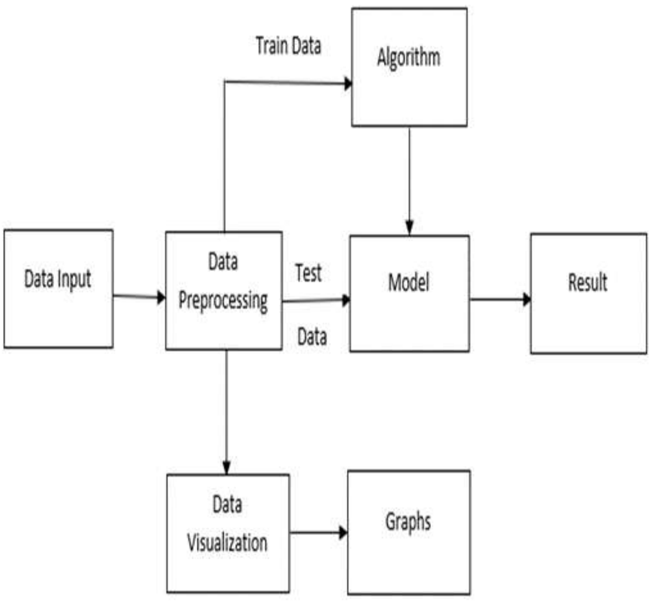
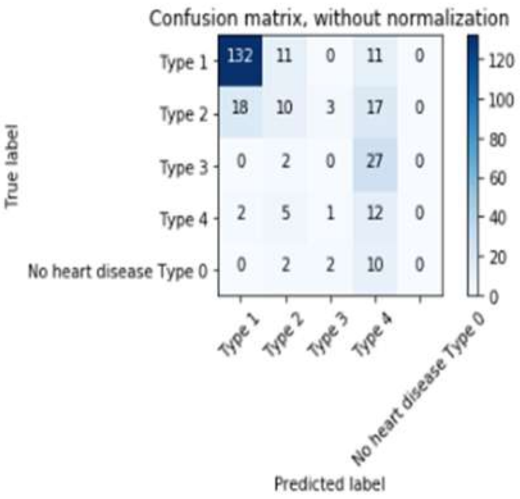


Figure 1: Block diagram of the system

Once the data is collected it is determined whether there is a missing value in any of the feature, and if present then that particular feature's value is taken as the mean value considering all the samples used in the dataset. This process is called data cleaning. After cleaning the dataset, it is divided into training set and testing set. Training set is used to train the algorithm and testing set is used for testing purpose. Data preprocessing in visualizing data in the form of graphs. Algorithm takes training dataset as the input to train on various samples. A total of four algorithms are used for evaluation namely random forest, naïve bayes, decision tree and linear regression. The data is fitted into these algorithms and the algorithm trains itself based on different instances in the dataset. Once the algorithm is set, it is tested using the remaining data kept for testing. The model predicts the accuracy and also predicts whether the patient is suffering from heart disease or not. If yes then predicts either type 1 or type 2 or type 3 or type 4 disease based on the dataset.

B. Implementation

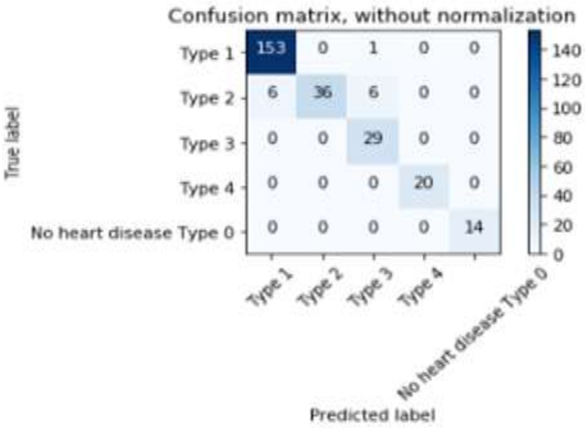
The implementation is carried out using four different machine learning algorithms, each having different accuracies. Linear regression is a linear model. Consider a model with linear relationship between single input variables [4] denoted as x and single output variable denoted as y, then y can be calculated from linear combinations of x. Fig 2 is confusion matrix that is used to describe performance of a classification model, on a particular test set where correct or true values are known.



	Type 1	Type 2	Type 3	Type 4	No heart disease Type 0
Type 1	132	11	0	11	0
Type 2	18	10	3	17	0
Type 3	0	2	0	27	0
Type 4	2	5	1	12	0
No heart disease Type 0	0	2	2	10	0

Figure 2: Confusion matrix for linear regression

It is easy to use machine learning algorithm, flexible. It is widely used algorithm because it can be used for both classification and regression tasks. Random forests algorithms are supervised learning algorithms.



	Type 1	Type 2	Type 3	Type 4	No heart disease Type 0
Type 1	153	0	1	0	0
Type 2	6	36	6	0	0
Type 3	0	0	29	0	0
Type 4	0	0	0	20	0
No heart disease Type 0	0	0	0	0	14

Figure 3: Confusion matrix for random forest

3. RESULTS

Different machine learning algorithms can be used for the classification process. Four different machine learning algorithms are used in this work, all of them give different accuracies. The algorithms used are linear regression, decision tree, naive bayes and random forest, the figure 4 below represents a bar graph of the accuracies. The accuracy scores of the algorithms are as follows:

- Random forest: 95.09%
- Decision tree: 91.32%
- Naive Bayes: 60.38%
- Linear regression: 58.11%

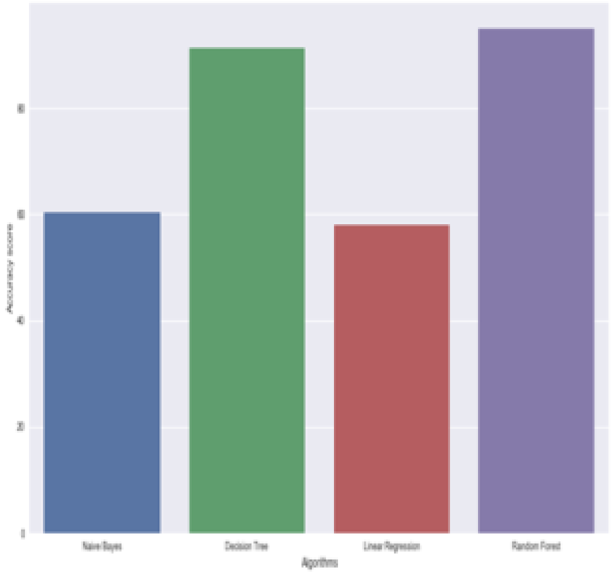


Figure 4: Bar graph of accuracy scores

4. CONCLUSION

CVD's have become one of the major concern due to the leading mortality rate. Major parts of the world are hit by cardio vascular diseases. With all these major factors prediction of cardiovascular diseases becomes very important and the models that predict CVD's would bring a major impact in reducing the mortality rate. Extended two-level classification is of more importance when compared to only two-level classification as the patients get a better overview of their current health regarding CVD scenarios. Prevention is always the basic step to stop any disease and hence that should be considered as a major concern and worked upon. Machine learning algorithms provides a suitable software environment to work on prediction and predicts the disease type accordingly. Heart disease prediction works best with random forest with highest accuracy of 95% as compared to the other three algorithms namely decision tree, naive bayes and linear regression.

REFERENCES

[1] "Chaitanya Suvarna, Abhishek Sali, Sakina Salmani","Efficient Heart Disease Prediction System Using Optimization Technique", https://ieeexplore.ieee.org/document/8282712

[2] "Ms. Tejaswini U.Mane", "Smart Disease Prediction System Using Improved K-Means and ID3 on Big Data", https://ieeexplore.ieee.org/document/8073517

[3] "Sushmita Manikandan", "Heart Attack Prediction System", ttps://ieeexplore.ieee.org/document/8389552

[4] "AH Chen, SY Huang, PS Hong, CH Cheng, EJ Lin", "HDPS: Heart Disease Prediction System", https://ieeexplore.ieee.org/abstract/document/820364 3