Journal of Engineering Sciences

Vol 11, Issue 5,May/ 2020

ISSN NO: 0377-9254

Chronic Health Disease Prediction Using Machine Learning

K.Surekha*,K.Pooja[1], K.M.S.Durgareddy[2], G.Roopavathi[3], M.Ramyasri[4], K.Rahul[5],

*Associate Professor, Dept of Computer Science & Engineering, VSM College Of Engineering,Ramachandrapuram.

[1,2,3,4] B.Tech Student, Dept of Computer Science & Engineering, VSM College Of Engineering, Ramachandrapuram.

Abstract: **In today's era everyone is trying to be conscious about health due to workload, one gives attention to the health when it shows any symptoms of some kind. But Chronic diseases like heart disease, kidney cancer, diabetes etc…, doesn't shows any symptoms at all or in some cases it doesn't show any disease specific symptoms. It is hard to predict the disease before it effects the person, But machine learning uses various techniques like KNN algorithm, Decision Tree, Super Vector Classification, Random Forest etc…, to predict the chronic disease before that effects the person. These Algorithms have the advantage of dealing with unstructured data and complex non-linear problems. In this paper we are predicting a sample chronic disease i.e.., heart disease using Logistic Regression to handle the unstructured data due to high accuracy of predicting data.**

Keyboards: **Heart Disease, Logistic Regression, Machine Learning Model, Data Set, Prediction.**

1. INTRODUCTION

The Chronic Health Disease Prediction is an end user application. It is an online web application that allows many users to predict the chance of getting a disease before it effects the person. The application is fed with various details of patient lab report for predicting according to the chronic disease i.e.., health disease. The application allows user to enter the health parameters for prediction. These parameters are different for different diseases. They need enter the correct values from the reports for accurate prediction rather than the assumption values. The blind entry of assumption values by the end user leads to false prediction.

The objectives of the "Chronic Disease Prediction Using Machine Learning" are:

a). Usually it's not possible to predict the chronic diseases before they effect's the person. Hence by using the Machine Learning, we create a model that helps us to predict the chance of getting a disease like Heart attack.

b). To decrease the risk of getting the sudden health attacks of chronic diseases.

c). Sometimes you or someone needs doctor help immediately, but not available the proper guidance for consultancy. Hence it also suggests the specialist along with hospital details.

Scope: It predicts the chance of getting chronic diseases for a person based on the lab report values. It only predicts whether the user condition is in an normal state or an abnormal state previously before the disease get attacks. But doesn't suggest in how many years or months the disease attacks the person, or if a person is already get attacked with a disease it doesn't predicts the stage of a disease.

2. EXISTING SYSTEM

Although the current system is manual and file based one. We realize that the system we are going to build that must give the solutions for wastage of time and space which affects the efficiency of daily activities performed at the hospital.

Everyone is a patient for sometime's and we all want good medical care. Assume that the doctors are medical experts and there exists good research behind every decision. But to predict the diseases before the person get affected by the disease. The doctor's needs in-depth research beyond the scope of a physician's work

Disadvantages: The current manual system has a lot of paper work and it doesn't deal with predicting the data previously. If someone wants to check the details of the available doctors the previous system doesn't provide any necessary detail of this type.

3. PROPOSED SYSTEM

In this project we have developed a ML system called online health prediction system using data analysis technique, which is used to simplify the task of predicting chronic diseases. While developing this system we combined the structured and unstructured data collected from the healthcare field which let us to assess the risk of disease before it effects.

Advantages:

1. To handle the structured and unstructured data we use k-means algorithm & linear regression to select the features, which thus used by the machine learning model to predict the disease.
2. User's can search for doctor's help at any point of time.
3. Doctors get more clients through online.

4. DATA OVERVIEW AND PREPROCESSING

The data set used in this is collected from the kaggle website, called Heart disease prediction dataset. It is an open dataset having 14 attributes of 304 different patients. But after preprocessing we use 11 attributes for further experiment as the 11 attributes are the most useful to predict the heart disease in a patient.

The complete description of the attributes is shown in below table:

age	Patient age in years
sex	Sex(1=male,0=female)
cp	Chest pain type --Value1: typical angina --Value2: atypical angina --Value3: non-anginal pain --Value4: asymptomatic
trestbps	resting blood pressure (in mm Hg on admission to the hospital)
chol	serum cholestoral in mg/dl
fbs	fasting blood sugar > 120 mg/dl 1 = true; 0 = false
restecg	resting electrocardiographic results -- Value 0: normal -- Value 1: having ST-T wave abnormality -- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria.
thalach	maximum heart rate achieved
exang	exercise induced angina 1 = yes; 0 = no
oldpeak	ST depression induced by exercise relative to rest.
thal	3 = normal; 6 = fixed defect; 7 = reversable defect.

Preprocessing: Data sets that are considered is taken from multiple sources which are normally not too reliable and that are too different in different formats, more than half of the time is spent in dealing with data quality issues when working on a ML problem statement. It is practically not possible to expect that the data will be perfect and in structured format. There may be problems in the data sets due to the human errors while collecting, limitation of measuring devices in solving problem statement, or faults in the data collection process. In this we remove missing values, null values, inconvenient values, duplicate values by box plot in machine learning which thus forms a structured data for further analysis.

i) Missing values: Initially the data that is used for training the model should be cleaned without having any null values. The null values are identified in the data sets as follows:

```
data1.apply(lambda x:sum(x.isnull()),axis=0)
```

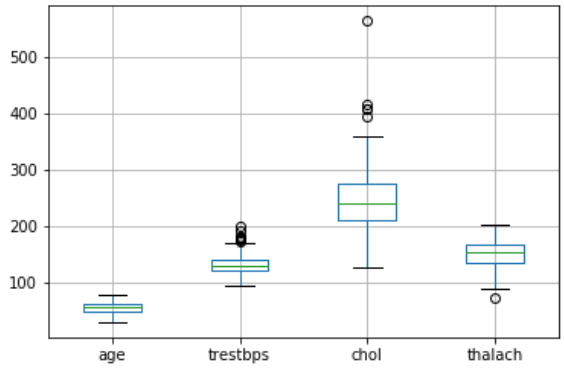
```
age         0
sex         0
cp          0
trestbps    0
chol        0
fbs         0
restecg     0
thalach     0
exang       0
oldpeak     0
ca          0
thal        0
target      0
dtype: int64
```

i) Outlier detection: Machine learning Algorithms are too complicated and more sensitive with the data sets for the statistical distribution of the input variables

that are simply called as a features. The data that are out of bounds are called Data outliers, that can spoil and mislead the training process of machine learning model resulting in last training times and with high training period with less accurate models which are lowers the result efficiency. Usually an Outlier is a rare chance of occurrence of a value into a classifier within a given data set. In Data Science (machine learning), data Outlier is an point that is observed too distant from other observations points in the data set. An Outlier may be due to misleading of data collection during the measurement of distance from other points or it may indicate experimental error which leads to less accuracy resultant.

Outliers are the points that are being the most extreme observations in the dataset, may include the maximum or minimum value in the dataset, or both, depending on the data set considered whether they are extremely high or low. Outliers are mainly detected using the box plot mechanism in machine learning. The outliers are identified by the quartile representation i.e.., the data points that are outside the quartiles are represented as the outliers in the dataset. Thus identified data outliers are converted into the frequently occurrence data point either by replacing the number by the absolute mean of all the points in the data set or by replacing using the standard deviation values of the three quartiles whose SD value is less than three etc.., These replacing is done by the following way:

Univariate method: The simplest
method for detecting the outliers in the data set is the use of box plots. Actually a box plot is a graphical representation of the display for describing the distribution of the data over all the data points in the data set. Box plots use the median and the lower, upper quartiles. The outliers are those that are far away from the median value i.e.., the central point. The maximum distance from the centre point that is to the allowed is usually a cleaning parameter. The test becomes sensitive when the value of cleaning parameter is too large.



The following figure determines the outliers present in the data set used for the model creation for predicting the heart disease. Thus detected outlier points are replaced by the respective mean of the data points in the data set considered or else by the standard deviation of the data points.

4. METHADOLOGY

Machine Learning: Now-a-days, max of all the people faces various chronic diseases due to the environmental condition and their living habits with these fast growing technology. So the prediction of those chronic disease at earlier stage or before occurring becomes important task. But the accurate prediction of a disease on the basis of symptoms and that too before effecting the person, becomes too difficult for a doctor. The correct prediction of disease in correct period is the most challenging task now-a-days. To overcome this problem data mining plays an important role in predicting the chronic disease like heart disease. Medical field has large amount of data growth as the years goes. Due to the increase amount of data growth in medical science and healthcare field it is more important for the accurate analysis on medical data which benefits large number of patients. I also helps in reducing the sudden health attacks. With the help of the disease data that is collected, machine learning model finds the required hidden pattern information in the huge amount of heart disease data that is gathered. We proposed general heart disease prediction based on lab reports of the patient.

- **How Machine Learns:** Usually a machine learning model may contain a mix of different

techniques and algorithms, the methods that are applied for learning or training a model can typically be categorized as three general types:

Supervised learning: The learning algorithm which contains the labeled data i.e.., features and then produces the desired output based on that features. For example, pictures of all the apples of different color are labeled "apple" in learning or training the model, then with the help of the algorithm the model identify the rules to classify pictures of apples of all the colors.

Unsupervised learning: The learning algorithm which contains the unlabeled data i.e.., it doesn't have any features while training a model and then produces the desired output based on that features.. For example, the e-commerce website like Flipkart, Amazon etc.., it displays similar items often bought together by the learning algorithm that is trained.

- **Reinforcement learning**: The learning algorithm that interacts with a dynamic and real time environment that learns from the previous output of the model and provides the feedback i.e.., output in terms of rewards and punishments. For example, self-driving cars being trained to not cross the Danger board.

Supervised learning algorithms apply the following techniques to describe the data:

1. **Classification** is the process where the new incoming data point is labeled based on past data samples used for model training and manually trains the algorithm, the model learns and recognize certain types of objects and categorize them accordingly based on it. The model has to know how to differentiate types of information, objects, perform an optical character, image recognition and categorizes the objects based on "yes" or "no" decisions.
2. **Regression** is a technique which is used for identifying the patterns formed by those data points and for calculating the predictions of the outcomes those are produced continuously. The model developed or trained by using regression need to understand the numbers, their values, and the grouping formed by them.

The most widely used supervised algorithms are:

a). Logistical Regression;
b). KNN algorithm;
c). Random Forest;
d). Support Vector Machines (SVM);
e). Decision Trees;

a). *Logistic Regression*: Logistic regression is one of the frequently used classification algorithm for assigning the observations i.e.., the new incoming data point to a set of classes which it suits. Some examples of the classification problems using logistic regression are Email spam or not spam, Online transactions Fraud or not Fraud, Tumor Malignant or Benign. The classification of the incoming points using Logistic regression is by the use of the logistic sigmoid function which transforms the output to return a probability value of it. This predictive algorithm for analysis is based on the probability concept using probability value. It is often called as a linear regression model but the main difference is that in the Logistic Regression is it deals with the cost function that is more complex, instead of a linear function this cost function can be defined as *Sigmoid function* or also known as the *logistic function*. Usually a Sigmoid function is a function used for mapping the real values in a dataset to the values in between 0 and 1. In machine learning, for mapping the predictions to the probabilities occurred we use sigmoid function. The conversion of all the data point values into the linear series between the 0 and 1 makes the task of classification easier.

b). KNN algorithm: KNN algorithms is another type of classification mechanism which use data for classifying the new entries based on similarities between them. For it we usually uses a distance function. Classification is done on the basis of major similarity with its neighbours. The new entry data point is assigned to the class with the less nearest neighbours with minimum distance. The increase in the number of nearest neighbours in a class, increases the value of k. It is used for solving both the classification and regression problems. It believes that the things with high similarity exists in close proximity. In other words, it forms the same cluster when the similar things are near to each other. KNN captures the idea of similarity, sometimes called distance, proximity, or closeness by calculating the distance between points

on a graph. KNN makes predictions using the training dataset that get splits while model creation.

For determining the new input point to a particular instance the distance measure is used. For the distance measure the most frequently used Euclidean algorithm is used for the real input values. We have different algorithms for calculating the distances between the points like Manhattan distance, Hamming distance, Minkowski distance. But most frequently used method is the Euclidean.

c). *Decision Tree*: It is most widely used algorithm for classification in a practical way. A decision tree is a flow of nodes from root to leaf i.e.., it contains flowchart like structure in which each internal node represents a test on a feature, each leaf node at the end of the flow chart represents the class labels i.e.., the decisions that are taken after computing the task and the branches of a tree represents the conjunction of features that lead to the class labels. Thus formed paths from root to leaf represent the classification rules. It follows an algorithmic approach for splitting of a data set using classification rules as a conditions. Decision Trees are a non-parametric supervised learning method mainly used for dealing both the classification and regression tasks. A decision tree is a tool that uses a tree-like model of all the decisions maked and their possible consequences.

d). *Random Forest:* Random forest is an approach which is used for dealing both the classification and regression tasks. But mainly used for classification problems rather than regression problems. Similar to the practical way, the forest is made up of large number of trees. Similarly, random forest algorithm creates multiple number of decision trees for the considered data samples and then gets the prediction from each of those decision trees and the prediction with the best solution is selected among all the prediction values. Random forest is simply termed as a group of decision trees which classifies the data points in a data set.

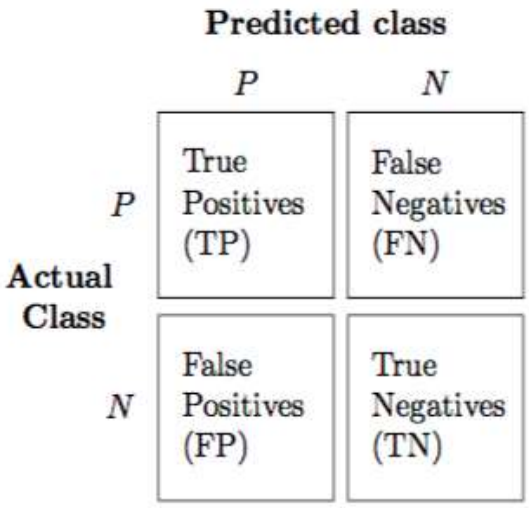
e). Support Vector Machines (SVM): SVM training algorithm is a rarely used classification technique which builds a model by assigning new incoming points to one category or the another which makes a non-probabilistic binary linear classification. It represents all the points in the space, so that all the points of different category are divided by a clear gap which is widely possible for performing a non-linear classification.

All these algorithms are applied for the model creation in prediction analysis. But the model get deployed only with the high accuracy values among all the different algorithms in predicting the disease. The accuracy of a model in machine learning is based on the splitting of data set that is considered for learning a model. The data is divided into a appropriate training and test sets with effective percentages and the model that is selected best suits foe the data for better accuracy. The deployment of a model with high accuracy acts as an interface between the model creation and prediction. Thus highest accuracy achieved while creating model gets deployed. The accuracy of models created by different algorithms is as follows:

Algorithms Used	Accuracy	Test data split	Train data split
KNN Classifier	48.275862	20%	80%
Logistic Regression	67.241379	20%	80%
Super Vector Classification	48.275862	20%	80%
Decision Tree	51.724138	20%	80%
Random Forest Classifier	81.034483	20%	80%

The Random forest Classifier with the accuracy of 81.034483 is the highest among all the algorithms used for creating a model. The test size used for all the algorithms is 20%, because the more the testing data the more will be the accuracy. Along with the Accuracy of the algorithm the confusion matrix thus produced should be considered. A confusion matrix is a 2X2 square matrix which is used to describe the performance of a model that is used for classification for a set of

test data which knows the true values. It makes the clear visualization of the performance of an algorithm. This denotes the count of correct and incorrect predictions by the model along with the count values. It contains two different rows and columns named true predictions and false predictions whose combination describes the performance.



The confusion matrices of different algorithms are as follows:

Algorithm	Confusion Matrix
KNN Classifier	[[12 18] [12 16]]
Logistic Regression	[[16 14] [5 23]]
Super Vector Classification	[[0 30] [0 28]]
Decision Tree	[[17 17] [11 17]]
Random Forest Classifier	[[18 8] [3 29]]

5. CONCLUSION

This project was successfully completed within the time span allotted. The project CHRONIC DISEASE PREDICTION USING MACHINE LEARNING gets implemented using the php, html, python, bootstrap. In this system the model is developed by various classification techniques but the model developed by using Random Forest Classifier with high accuracy get deployed which helps in accurate prediction of chronic disease. The system had been developed in an attractive dialogs for user interface. So user with the basic knowledge can also operate and use the system easily. The speed and accuracy of the system is maintained in proper way by using structured and unstructured data.

7. REFERENCES

- Diabetes Prediction Using Machine Learning Techniques
  Author: Tejas N. Joshi,2018
- Implementation of Machine learning model to predict Heart Failure Disease
  Author: Fahd Saleh Alotaibi,2019
- Heart Disease Prediction Using Machine Learning Techniques
  Author: V.V.Ramalingam, M.Karthik Raja, 2018
- Prediction of Chronic Kidney Disease Using Machine Learning Algorithm.
  Author: Siddheshwar Tekale, 2018
- Heart Disease Prediction Using Machine Learning
  Author: Priya male wadkar, Omkar Baswat, 2019
- Heart Disease Prediction Using Logistic Regression Algorithm Using Machine Learning
  Author: Reddy prasad, N.Deepa, 2019
- OnSet Diabetes Prediction Using Machine Learning Techniques
  Author: M.D.Aminul Islam, Nusrat Jahan, 2017

- Statistical Analysis and Predicting Kidney Disease Using Machine Learning Algorithms
  Author: P.Swathi Baby, T.Pandurunga vital, 2015

- Detection of Chronic Kidney Disease Using Machine Learning Algorithms With Least Number Of Predictions.

  Author: Marwa Almasoud, Tomas E Ward, 2019