International Journal of Engineering Applied Sciences and Technology, 2020
Vol. 5, Issue 1, ISSN No. 2455-2143, Pages 174-179
Published Online May 2020 in IJEAST (http://www.ijeast.com)

DISEASE PREDICTOR: ALGORITHMS TO REDUCE THE FAULTY PREDICTIONS

Nikhil Sharma, Mimansha Singh, Yash Gupta
Student IMS Engineering College
Ghaziabad, India

Abstract -- There are many tools related to disease prediction, but particularly for any specific disease such as heart, diabetes or cancer. But generally there is no such tool that is used for prediction of other different diseases. So Disease Predictor helps for the prediction of several other diseases making use of different machine learning algorithms to reduce the faulty predictions.

Keywords— **Naïve Bayes, Random Forest, Extra Tree, SVM.**

I. INTRODUCTION

Disease Predictor is a tool based on some good machine learning algorithms whose main function is to diagnose the diseases of patients with which a patient may be infected.

According to a report by the World Health Organization (WHO), more than 138 million patients are harmed every year by doctors' errors among which 2.6 million mistakes resulting the death of people.

Actually, there are many tools related to disease prediction, but particularly for any specific disease such as heart, diabetes or cancer "Krishnaiah V.,Theresa Princy R. et.el (2016,2016,2005,2012,2014) emphasize the same in their study". But generally there is no such tool that is used for prediction of other many different diseases. Also this tool helps to find accuracy in the cases where the symptoms of any 2 diseases are not differentiable and causes prediction fault by doctors resulting in some serious conditions. So Disease Predictor helps for the prediction of several other diseases making use of different machine learning algorithms to reduce the faulty predictions and increase the accuracy.

This project aims to provide a platform to predict the occurrences of diseases on the basis of various symptoms in which user can select various symptoms and can find the diseases with their probabilistic figures. Here we have used four predefined machine learning algorithms that are:

- Naïve Bayes
- Extra Tree
- Random Forest
- Support Vector Machine (SVM)

II. LITERATURE SURVEY

Numerous studies and researches have been done that focus on the diagnosis of diseases. They have used different kinds of machine learning algorithms and other strategies but they lag in diagnosing more than one disease and also they use only a single algorithm which cannot give an accurate result always. Therefore we proposed a tool for diagnosing more number of diseases with greater accuracy by use of multiple machine learning algorithms.

III. DISCUSSION

A. DATASETS:

Taking into account the importance of datasets and their impact in the obtained result, it is very crucial to discuss datasets used in this study. We found it really difficult to find any dataset which includes diseases and their symptoms and was sufficient for the training of the model as we know that if we train our model with less data then we cannot be ensured that our model is predicting fine and with optimum accuracy. There are many datasets which include diseases along with their symptoms but the problem with them is, that they are pretty small for the training purpose. So we have merged different datasets from many resources so that we could ensure that training is fine enough for predicting any disease on the basis of symptoms.

B. DISEASE PREDICTION TECHNIQUES:-

There are many Machine Learning algorithms available for use but the problem occurs when we have to make a choice among those algorithms. Many studies shows that for the classification problems where we have more than 2 or 3 classes, algorithms such as Linear Regression, Logistic regressions and KNN do not perform well. While the algorithms like Decision Tree, Random Forest, Naive Bayes and Support Vector machine are among some algorithms which are accepted as some of the best classification algorithms "Lakshmi B.N et.el (2015) emphasize the same in their study".Instead of only using

International Journal of Engineering Applied Sciences and Technology, 2020
Vol. 5, Issue 1, ISSN No. 2455-2143, Pages 174-179
Published Online May 2020 in IJEAST (http://www.ijeast.com)

Decision Tree it is also seen that combination of Decision Tree and Random Forest performs very well in such scenarios "Kaur Beant et.el(2014) emphasize the same in their study".

So, this study also uses different machine learning algorithms for the prediction of the diseases. The other thing which is used is the Graphical User Interface (GUI) made with the help of the tkinter library of Python.

Naïve Bayes:

Naive Bayes is a classification algorithm which follows the bayes rule. It is basically a probabilistic model which produces the probability of happening of any particular task.

$$P(A|B)=\frac{P(A)P(A)}{P(B)}$$

Figure (1): Naïve Bayes Formula

So, by using this classifier we are able to find out the probability of happening of any event say A when probability of happening of any other say B is known to us. In this probabilistic model we make an assumption that the events/features which we are considering are independent of each other. In other words we can say that one particular feature/event does not affect the other. That is why it is known as Naive.

Extra Tree:

Extra Tree is an abbreviation for Extremely Randomized Tree. Extra tree is nothing but an advance version of Random Forest. The extra tree classifier is said to be an advance version by the fact, that it is highly randomized.

The Extra classifier also make use of a forest which consist of many decision trees. In this forest each tree makes its prediction. These trees are highly DE-co related which provides model an ability to predict the best result.

The difference between extra tree and Random forest classifier lies in their tree construction method. An extra classifier make use of whole dataset for constructing the tree while random forest make use of bootstrapping for its tree construction.

Extra tree classifier predicts the result in a very less time because it do all the things by randomization procedure. The splitting in extra tree is also a random selection from the chosen features. It make use of gini index widely for constructing the tree for learning process.

$$Entropy(S)= \sum_{i=1}^{c} -p_i \log_2 p_i$$

$$Gain(S,A)=Entropy(S) - \sum_{veValues(A)} \frac{Sv}{S} Entropy(Sv)$$

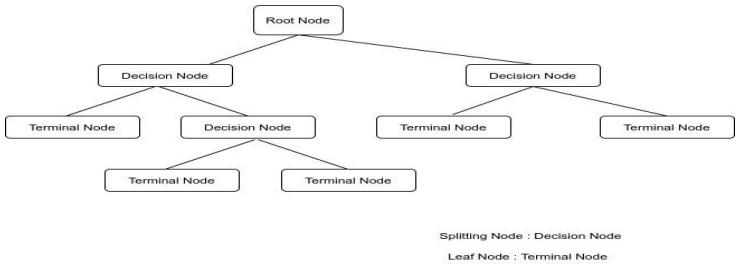


Figure (2): Tree Representation

Random Forest:

Random forest, as its name suggests, it is the collection of a large number of the individual decision trees that operate as an ensemble. In a random forest each individual decision tree predicts a class as its output and that class which have maximum number of votes becomes the output for the whole model.

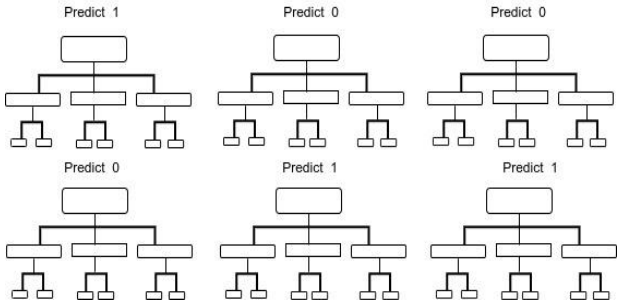


Figure (3): Different Decision Trees

In random forest algorithm having a low co-relation among the various decision tree plays a very important role as these trees do not affect each other and provides a kind of a reinforcement for the other tree. Due to this reason of low co-relation we know that they can produce the best result. The reason behind this effect is that these trees protect each other from their individual error and provides a kind of reinforcement as well. We cannot ignore the possibility that some of the tress may provide wrong output as well but as whole group they are able to move in the right direction due to reinforcement and votes. There are some prerequisite which makes the process of Random forest to work well. These prerequisite can be defined as following:

International Journal of Engineering Applied Sciences and Technology, 2020
Vol. 5, Issue 1, ISSN No. 2455-2143, Pages 174-179
Published Online May 2020 in IJEAST (http://www.ijeast.com)

1. The features of data points should have real a signal so that the model can guess the actual output instead of some random guessing.
2. The output produced by the each tree should have a low co-relation with others so that the training can be headed in the right direction.

Support Vector Machine:
SVM which an abbreviation for Support Vector Machine is one of the best classification algorithm. The main principle of support vector machine is to find a best fit hyperplane which can easily distinguish among different classes.

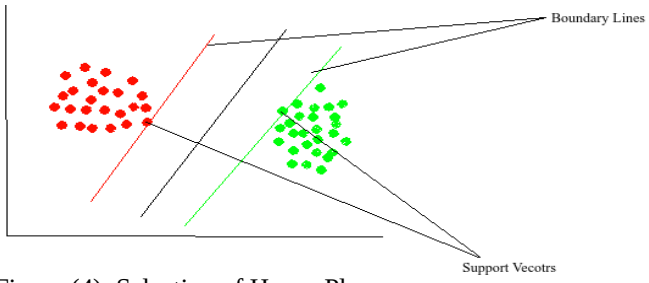


Figure (4): Selection of Hyper Plane

There are many ways by which we can find out the hyperplanes but we have to make a choice among those. This is one most important part of whole algorithm. To select any particular hyperplane we take helps of vectors. These vectors are also features but these features are those which lies of the plane which distinguishes classes.

We consider the best plane which has maximum margin i.e. the plane which has maximum distance among data points of two different classes will be considered as best plane. We do so because this helps the methodology by providing some kind of reinforcement so that other data points can be classified with more confidence.

These hyperplanes are the actual boundary lines which helps in classifying the data points. Support vector machine supports multi-dimensional planes. The dimension of any plane depends upon the number of feature used.

The procedure or the functions through which the planes are identified are called kernels. There are many kernels which are used. Some of the popular kernels are as-

1. Linear Kernel
2. RBF kernel
3. Gaussian Kernel

So for this study we are making use of RBF kernels. RBF kernels make use of Radial basis functions for their implementation in order to find out the plane.

If we have two samples named x and x', represented as feature vectors in input space, then RBF kernel can be defined as-

$$K(X,X') = \exp(-\frac{||X-X'||^2}{(2\sigma^2)})$$

Where $||X - X'||^2$ is recognized as squared Euclidean distance between the two feature vectors. σ is a free parameter.

The above equation can also be written by including a special parameter: $\gamma = \frac{1}{2\sigma^2}$

So, the equation becomes:

$$K(X,X') = \exp(-\gamma||X - X'||^2)$$

IV. CASE STUDY

A. Data Collection

In this study, the dataset of diseases and their symptoms is used to train the model which is first divided into 2 different datasets which are training data and test data.

Our dataset includes 3632 data rows and 133 columns that are features in the training dataset while 1329 rows and 133 features in the testing dataset.

Features are basically the name of the symptoms so we have included 132 symptoms and 40 different types of diseases in our dataset which we have collected from different datasets which are used in some of the researches.

B. Data Pre-Processing

We have applied techniques such that the values of the features were normalized to range between 0 to 1. Then, data was standardized to have a mean of 0 and standard deviation of 1.After applying the pre-processing it becomes ready for training. Then the dataset is split into a proportion of **70:30** where 70% of data is used to train the model while 30% data is used to test the model.

C. Implementation

There are two phases of the project. These can be divided as:

International Journal of Engineering Applied Sciences and Technology, 2020
Vol. 5, Issue 1, ISSN No. 2455-2143, Pages 174-179
Published Online May 2020 in IJEAST (http://www.ijeast.com)

Model Training
We have used the algorithms which perform good in the classification of data because all we need is to classify the diseases on the basis of the symptoms of the diseases. After making the classifier's objects we have fed the training set into the model and trained our model. After the training we have tested our data with our testing set.

Using Trained Model for the detection of the disease

To use the trained model, a Graphical User Interface (GUI) is made using the Tkinter library of Python which looks like:

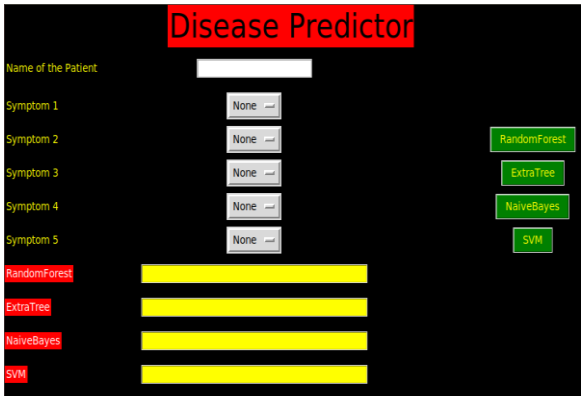


Figure (7): GUI of the tool

Here we have given a 5 drop down selection menu for the selection of the symptoms and 4 buttons for the disease prediction (each using a different algorithm).

D. Result:

On the basis of the data trained in our model we are getting different accuracies for different algorithms implemented which are as:

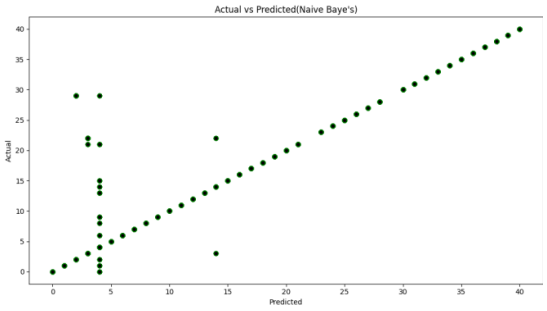
S.NO	Classifier	RMS-Error	Accuracy
1	Random Forest	0.610	91.3%
2	Extra Tree	0.515	93.6%
3	Naive Bayes	0.531	92.7%
4	SVM	0.518	93.4%

Figure (6): Accuracy and RMS error

We have also plotted some graphs to show the comparative results of actual value to the predicted value for each algorithm:

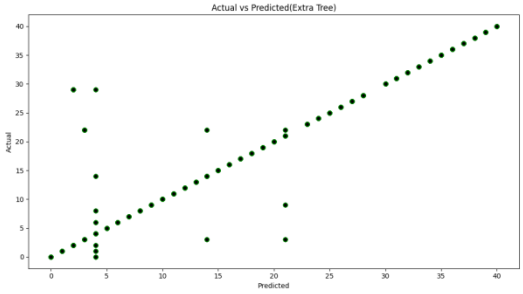
Naïve Bayes :

The accuracy score for Naive Bayes algorithm was found 0.92. Given below is the graph between the actual disease and the predicted disease by using Naive Bayes algorithm.



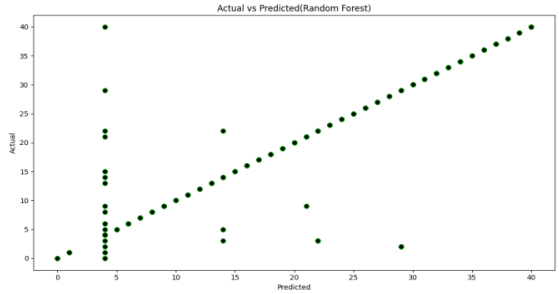
Extra Tree:

The accuracy score for Extra Tree algorithm was found 0.936. Given below is the graph between the actual disease and the predicted disease by using Extra Tree algorithm.



Random Forest:

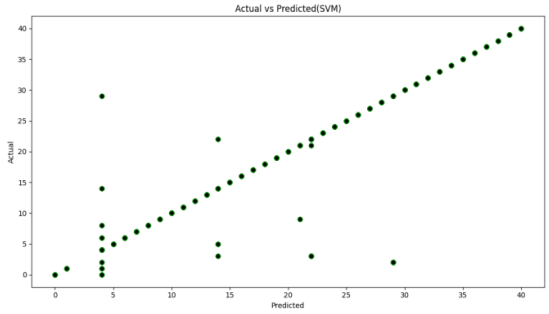
The accuracy score for Random Forest algorithm was found 0.913. Given below is the graph between the actual disease and the predicted disease by using Random Forest algorithm.



International Journal of Engineering Applied Sciences and Technology, 2020
Vol. 5, Issue 1, ISSN No. 2455-2143, Pages 174-179
Published Online May 2020 in IJEAST (http://www.ijeast.com)

Support Vector Machine:

The accuracy score for Support Vector machine algorithm was found 0.934. Given below is the graph between the actual disease and the predicted disease by using Support Vector Machine algorithm.



V. FUTURE WORK

- This project has not implemented recommendation of medications to the user. So, medication recommendation can be implemented in the project.
- History about the disease for a user can be kept as a log and recommendation can be implemented for medications.
- Dataset can be increased with more diseases and their symptoms to further improve the accuracy.
- This study does not involves the classification of diseases on the basis of age. So this feature may provide better results.

VI. CONCLUSION

Researchers are passionate to try different types of classifiers and build new models with an effort to enhance the accuracy of the model they use. We have also tried to enhance the accuracy of our system so that we can avoid the faulty prediction of diseases.

Overall an accuracy of around 92% was found in this model which is far better than other existing models.

VII. REFERENCES

[1] Krishnaiah V., February 2016, Heart Disease Prediction System using Data Mining Techniques and Intelligent Fuzzy Approach: A Review International Journal of Computer Applications (0975 – 8887) Volume 136 – No.2

[2] Theresa Princy R. and Thomas J., 2016, Human heart disease prediction system using data mining techniques, DOI: 10.1109/ICCPCT.2016.7530265.

[3] Delen, D., Walker, G., & Kadam, A., 2005, Predicting breast cancer survivability: A comparison of three data mining methods. Artificial Intelligence in Medicine, 113-127.

[4] Gomathi, K., July2012, An empirical study on breast cancer using data mining techniques. International Journal of Research in Computer Application & Management, 97-102.

[5] Kumara, M., Vohra, R., Arora, A., 2014, Prediction of diabetes using Bayesian network. International Journal of Computer Science and Information Technologies, 5174-5178.

[6] Chaitrali S., Dangare and S. Apte Sulabha, June 2012, Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques, International Journal of Computer Applications (0975-888), vol. 47, no. 10, pp. 44-48.

[7] Obenshain M.K, 2004, Application of Data Mining Techniques to Healthcare Data, Infection Control and Hospital Epidemiology, 25(8), 690-695.

[8] Lakshmi B.N. ,Indumathi T.S. and N. Ravi, 2015, A comparative study of classification algorithms for predicting gestational risks in pregnant women, International Conference on Computers, Communications, and Systems (ICCCS), Kanyakumari,, pp. 42-46, doi: 10.1109/CCOMS.2015.7562849.

[9] Kaur Beant and Singh H. Williamjeet, October 2014, Review on Heart Disease Prediction System using Data Mining Techniques, *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 2, no. 10, pp. 3003-08.

[10] Jyothi Thomas and Kulanthaivel G., 2013, Preterm Birth Prediction Using Cuckoo Search Based Fuzzy Min-Max Neural Network, *International Review on Computer and Software (IRECOS)*, vol. 8, no. 8, pp. 1854-62.

[11] H. Blockeel and J. Struyf, 2002, Efficient algorithms for decision tree cross-validation, Journal of Machine Learning Research, vol. 3, pp. 621-650

[12] R. Duriqi, V. Raca and B. Cico, 2016, Comparative analysis of classification algorithms on three different datasets using WEKA", 5th Mediterranean Conference on Embedded Computing (MECO), pp. 335-338, DOI: 0.1109/MECO.2016.7525775.

[13] Syed Umar Amin, Kavita Agarwal, Rizwan Beg, 2013, Genetic neural network based data mining in prediction of heart disease using risk factors, DOI**:** 10.1109/CICT.2013.6558288.

[14] Ektaa Meshram, Gajanan Patle, Dhiraj Dahiwade, 2019, Designing Disease Prediction Model Using Machine Learning Approach, DOI: 10.1109/ICCMC.2019.8819782.