IDENTIFICATION OF DISEASE PREDICTION BASED ON SYMPTOMS USING MACHINE LEARNING

[1]J.Sravanthi, [2]D.Srisha, [3]Sravani Reddy, [4]I.Madhurima
[1,2,3]Assistant Professor, [4]UG Student, [1,2,3,4]Department of CSM, Kasireddy Narayanareddy College of Engineering and Research, Hyderabad, Telangana

ABSTRACT-
The development and exploitation of several prominent Data mining techniques in numerous real-world application areas has led to the utilization of such techniques in machine learning environments, in order to extract useful pieces of information of the specified data in healthcare hospitals and free service centers provided by doctors etc. The analysis will be accurate of medical database benefits in early disease prediction, patient care. The techniques of machine learning have been successfully employed for applications including Disease prediction. The aim of developing classifier system using machine learning algorithms is to immensely help to solve the health-related issues by assisting the doctors/ directly patients to predict and diagnose diseases at an early stage. A Sample data of 4920 patients' records diagnosed with 41 diseases was selected for analysis. A dependent variable was composed of 41 diseases. 95 of 132 independent variables (symptoms) closely related to diseases were selected and try to minimize those symptoms to related categories. This research work carried out demonstrates the disease prediction system developed using Machine learning algorithms such as Decision Tree classifier, Random forest classifier, and Naïve Bayes classifier, SVM, KNN.

Keywords: Machine Learning, Data mining, Decision Tree classifier, Random forest classifier, Naive Bayes classifier, SVM, KNN.

INTRODUCTION

Health information needs are also changing the information seeking behaviour and can be observed around the globe. Challenges faced by many people are looking online for health information regarding diseases, diagnoses and different treatments that will take lot of time and waste of money. If a recommendation system like a predictor can be made for doctors and medicine while using review mining will save a lot of time. In this type of system, the user interface problem in understanding the difficult medical vocabulary such as scientific names. User is confused because a large amount of medical information on different types of symptoms are available. The idea behind this system is to adapt to cope with the special requirements of the health domain related with users.

With the rise in number of patient and disease every year medical system is overloaded and with time have become overpriced in many countries. Most of the disease involves a consultation with doctors to get treated. With sufficient data prediction of disease by an algorithm can be very easy and cheap. Prediction of disease by looking at the symptoms is an integral part of treatment. In our project we have tried accurately predict a disease by looking at the symptoms of the patient. [2]We have used 5 different machine learning algorithms for this purpose and gained an accuracy of 92-96%. Such a system can have a very large potential in medical treatment of the future. We have also designed an interactive interface (GUI) to facilitate interaction with the system. We have also attempted to show and visualized the result of our study and this project.

As the use of internet is growing every day, people are always trying to know the new things of curious to know different new things. People always try to refer to the internet if any problem occurs. People do not have immediate option when they suffer with particular disease they have to wait for lot of time for tests. So, this system can be helpful to the people as they have access to internet 24 hours and easily find the medication on time.

Mainly for completing this we have to focus on the past data that contains the various different types of diseases of various types of patients will be suffering from the past several years. The main aim of developing this classifier system using machine learning algorithms is to immensely help to solve the health-related issues by assisting the physicians to predict and diagnose diseases at an early stage. A Sample data of 4920 patients' records diagnosed with 41 diseases was selected for analysis. A dependent variable was composed of 41 diseases. 95 of 132 independent variables (symptoms) closely related to diseases were selected and categorized into certain groups. This research work carried out demonstrates the disease prediction system developed using Machine learning algorithms such as Decision Tree, Random forest, Naïve Bayes, Support Vector Machine, and K-Nearest Neighbor algorithms.

PROPOSED ALGORITHM

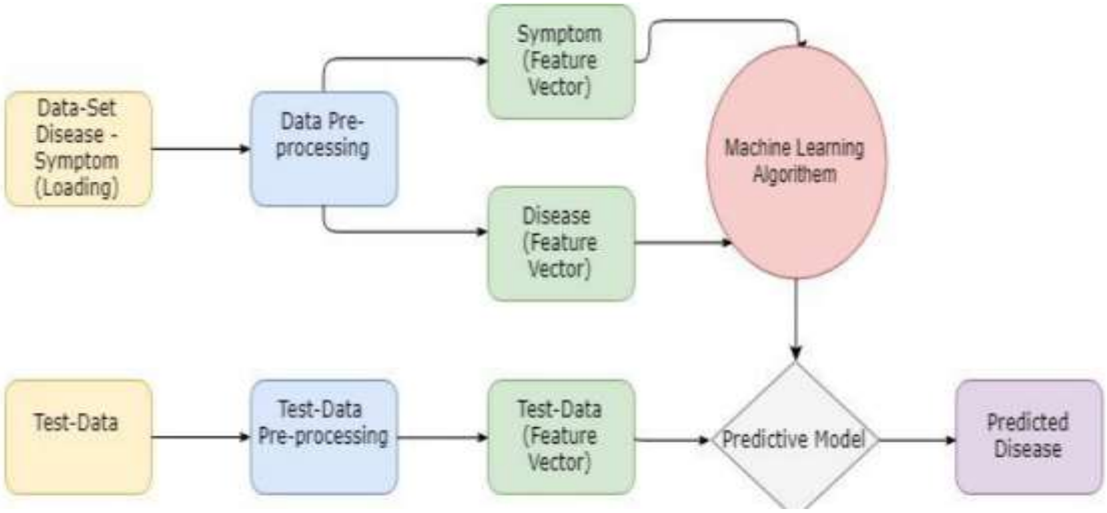


Figure 1. Methodology of algorithm

Classification and regression techniques are an essential part of machine learning and data mining applications nowadays. Approximately from the past data 70% of problems is in Data Science are classification problems. There are lots of classification and regression problems that are available, but the logistic regression is common and is a useful and accurate regression method for solving the binary classification problem. Another category of classification is Multinomial classification, in which it handles the issues where multiple classes are present in the targetvariable.

The data was sourced from Kaggle website as disease prediction using symptoms containing various types of symptoms as 0's and 1's and final attribute will be the type of disease. The data consists of 4920 records with 132 attributes. We used this dataset to predict the disease whether the patient will be suffering based on symptoms the patient will have. The dataset consists of 4920 rows and 132 columns. The dataset contains several numerical columns providing various information on symptoms for different diseases details. This dataset doesn't contains any missing values. So we have to easily implement the dataset without removing any records from the dataset.

The data collected consists of about 4920 records and 132 attributes and target attribute that is prognosis such as types of diseases. To predict the type of disease that a patient will be suffering we need to build a machine learning model. For that we have used the Jupyter notebook and IBM cloud for training the model and deploying it. Building a machine learning includes the following steps.

- Data Preprocessing
- Feature Extraction
- Model Training
- Prediction
- Deployment toPrediction

Data Preprocessing:

Data preprocessing includes 5 steps they are:

- Importing Libraries and Reading theDataset
- Check the missing values and considering the Correlational heatmap.
- Separating the independent and dependentvariables.
- Converting the data into numpy array and perform Encoding on categoricalvariables.
- Splitting the dataset for training andtesting.

We have splited the the dataset as 60% of the data to train the model and 40% to test the model. For this we need to import the "train_test_split" from sklearn package. In this splitting we use class which have attributes like test_size that specifies the percentage of test data and random_state that can have values 0 or 1 which is used to set the test data from the dataset.

[5]Here we have to try with different types of machine learning algorithms like Navie Bayes, Support Vector Machine, KNearestNeighbour, Random Forest, Decision Tree algorithms. We have to implement these different types of algorithms in order to find the best accuracy by using the train and test data from the previous dataset. By using these different 5 algorithms we have to find the best accuracy and try to model the data by using that best accurate algorithm.

Naive Bayes Algorithm

Naive Bayesian algorithm is a simple technique used for the purpose of classification. The workflow of the algorithm is based on the probabilistic method. This method includes strong independent assumptions. So it is well known as probabilistic classifier. It provides the feature to construct the classifier models that assign class labels to problem instances.

The general formula for calculating the conditional probability is

- $P(H|E) = (P(E|H) * P(H))/P(E)$
- $P(H)$ = probability of Hypothesis H being true.
- $P(E)$ = probability of Evidence
- $P(E|H)$ = probability of Evidence given that Hypothesis is true
- $P(H|E)$ = probability of Hypothesis given that Evidence is present.

Work Flow of Naive Bayesian

1. First, collect the dataset and split it into two datasets namely training dataset and testing dataset.
2. Second, perform the training model on Training dataset.
3. Take the input from the training dataset.
4. Let the model classify and make prediction based on given inputs.
5. Perform Normal distribution on the dataset to calculate the accuracy.

Decision Tree Algorithm

Decision tree algorithm is one of the Machine Learning Algorithm. It falls under the category of supervised learning techniques. It is used for the analysis of classification. However, it is also useful in regression analysis. Decision tree uses a set of tree like models of decisions and its possible consequences to make the decisions in an optimized manner. These decisions are to be resolved by using conditional control statements. Thus, in simple terms, we can conclude that decision tree algorithm is a set of trees that include bunch of conditions to generate a model of data at every node of a tree.

Work Flow of Decision Tree

The following steps are to be considered in the decision tree algorithm.

1. Consider we have taken a dataset that includes the features and target attribute.
2. Now give the training dataset to an algorithm, i.e. decision tree algorithm, it will generate a model that is used for classification and prediction. This is how it will create tree-structured classifier.
3. Now, the model will take the dataset as an input and based on what mechanisms says or whatever the rules defined in the model it will provide a class or target attribute, which will tell us that given input belongs to a particular class or target attribute.
4. Here, the decision tree performs its operations in an if-else conditional manner as it consists of a lot of decision trees and our task is to find the best solution for the given input.

Random Forest Algorithm

Random forest is a type of algorithm which is used for classification and regression. As we know the random forest is a one of supervised learning algorithm which means random forest algorithm uses the technique based on supervised learning if we talk about supervised learning in simple word supervised means a supervisor which gives the instruction, i.e. training data which gives input and output and based on the input and output of training data we are going to prepare a model and we will give new input to that model and check the output whether the valid output is coming.

The random forest is a type of ensemble classifier which is using the decision tree algorithm in a randomized fashion. It consists of many trees which are calleddecision trees and these trees are of different structures and to make a decision tree we are choosing features and samples randomly from the training dataset and that’s how we construct many decision trees and combined all the decision trees makes a random forest. Here, we are going to explain how a random forest works:

1. Initially, we should have training data which consist of various attributes and target attribute.
2. Now, we have to make a decision tree, to make a decision tree we have to generate BD (Bootstrap dataset) and to make BD we have to do sampling which means we have to pick any sample randomly from the training dataset and put it into Bootstrap dataset. Duplication is allowed with less frequency.
3. Using BD we have to plot a decision tree in a randomized fashion and calculate how we can choose the root node from the BD which is producing the best split of samples.
4. Again do the splitting of features for child node and provide the leaf node to the child node after splitting of features.
5. Repeat steps 2–4 and makes as many decision trees as we can.
6. Take the test tuple and let the model classify and predict the output of the given test tuple.
7. Now, calculate the votes produced by various decision trees.
8. Consider the majority of votes produced for the target attribute of test tuple and that will be a final prediction.

Support Vector Machine

Workflow of the Support Vector Machine (SVM) algorithm implementation as follows.

1. Vectorization step for each project in the database.
2. Learning step using the Support Vector Machine (SVM) algorithm with the training sets.
3. Prediction of project classification.

K-Nearest Neighbor

1. Load the data.
2. Initialize K to your chosen number of neighbors(n value) n=1.
3. For each example in the data

3.1 Calculate the distance between the query and the current from the data.

3.2 Add the distance and the index to an ordered collection

4. Sort the ordered collection of distances and indices from smallest to largest (in ascending order) by the distances from one to another.
5. Pick the first K entries from the sorted collection of data
6. Getting the labels of the selected K entries.
7. If regression, returns the K labels mean.
8. If classification, returns the K labels mode.

EXPERIMENT ANDRESULTS

User Interface before entering the symptoms.

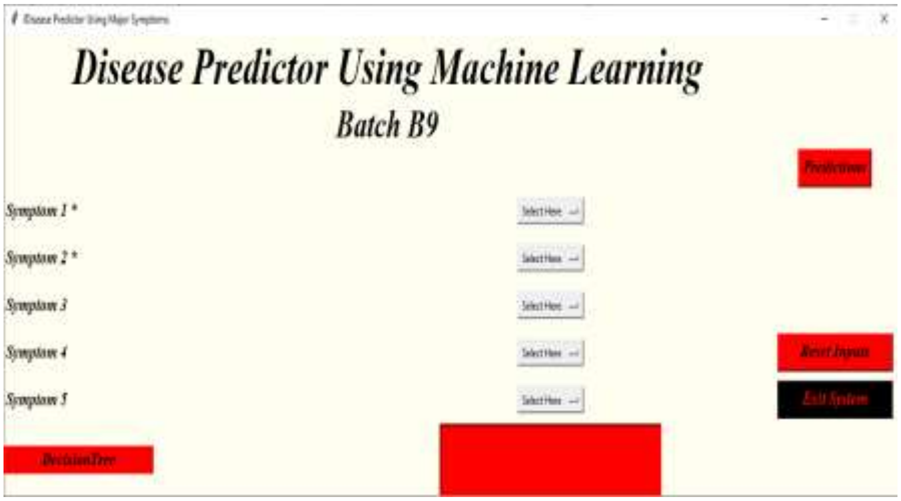
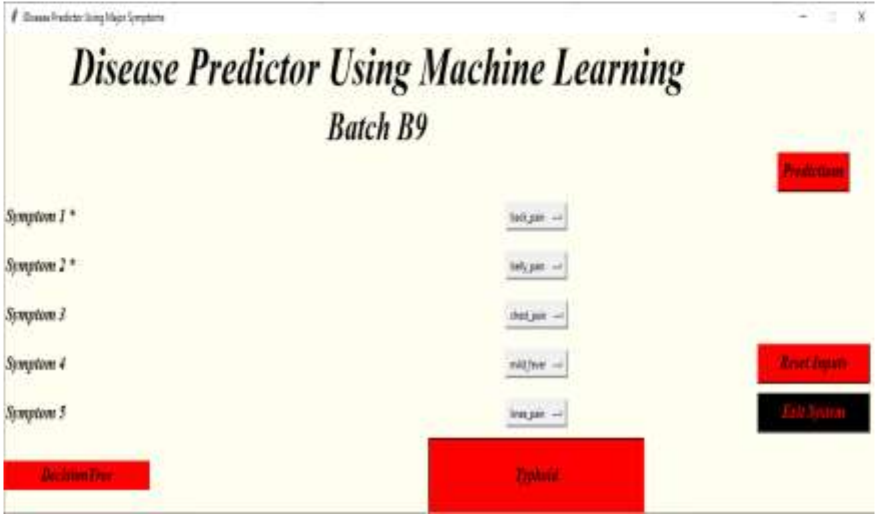


Figure 2. User Interface before entering symptoms

After entering the symptoms and predicting the type of disease.



Disease Predictor Using Major Symptoms

Disease Predictor Using Machine Learning

Batch B9

Symptom 1 * back_pain

Symptom 2 * belly_pain

Symptom 3 chest_pain

Symptom 4 mild_fever

Symptom 5 knee_pain

DecisionTree Typhoid

Figure 3. User Interface after entering the symptoms.

Summary/ table view of all algorithms

Table 1. Experiment results

ALGORITHM	ACCURACY
Random Forest	94.12
KNN	92.68
Naive Bayes	94.25
SVM	95.01
Decision Tree	95.13

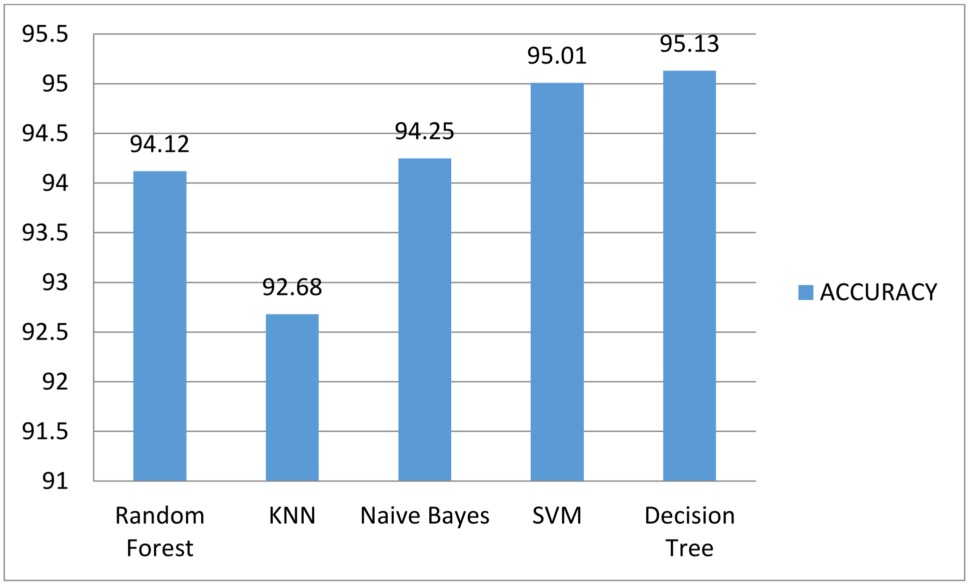
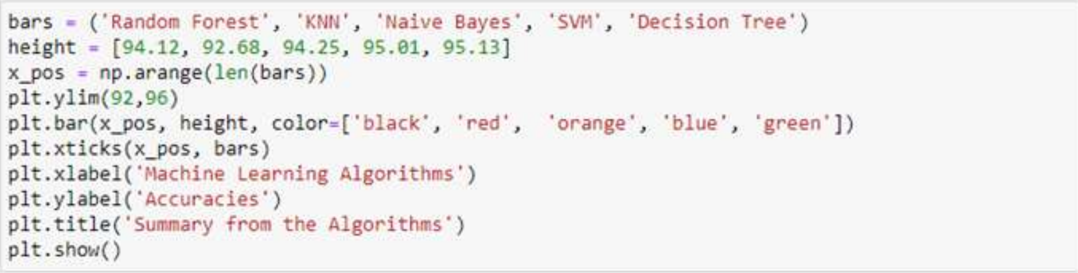


Figure.4. Comparison of accuracy levels of various models



```
bars = ('Random Forest', 'KNN', 'Naive Bayes', 'SVM', 'Decision Tree')
height = [94.12, 92.68, 94.25, 95.01, 95.13]
x_pos = np.arange(len(bars))
plt.ylim(92,96)
plt.bar(x_pos, height, color=['black', 'red',  'orange', 'blue', 'green'])
plt.xticks(x_pos, bars)
plt.xlabel('Machine Learning Algorithms')
plt.ylabel('Accuracies')
plt.title('Summary from the Algorithms')
plt.show()
```

Figure.4. Summary of algorithms

Now the accuracy score is predicted using both the predicted data(y_predict) and the original data(y_test). On caluclating we got accuracy score about 95% which is a good accuracy rate.

Discussions

Here we have to use Decision Tree Classification algorithm comparing with 4 different algorithms. Finally Decision Tree will get the best accuracy of 95.13. Then we have to implement this algorithm in-order to predict the type of disease.

Finally we have to develop an GUI that will helps to predict the type of disease by using major symptoms by a patient will be suffering at that time.

CONCLUSION

We set out to create a system which can predict disease on the basis of symptoms given to it. Such a system can decrease the rush at OPDs of hospitals and reduce the workload on medical staff. We were successful in creating such a system and use decision tree algorithm to do so. On an average we achieved accuracy of ~95%. Such a system can be largely reliable to do the job. Our system also has an easy to use interface. It also has various visual representation of data collected and results achieved.

Machine learning can be used for a variety of tasks. In this article, we used a machine learning algorithm to predict the disease based on the major symptoms. We also performed exploratory data analysis to find interesting trends from the dataset. For the sake of practice, I will suggest that you try to predict the disease that the patient is more likely by suffering, depending upon the symptoms.

Machine Learning can be a Supervised or Unsupervised. If you have lesser amount of data and clearly labelled data for training, then you have to go for Supervised Learning. Unsupervised Learning would give better performance, accurately outputs and results for large data sets. If you have a huge data set easily available, go for deep learning techniques instead of machine learning. You also have learned about Reinforcement Learning.

REFERENCES

[1] Mir A, Dhage SN (2018) Diabetes disease prediction using machine learning on big data of healthcare and domain system. In: 2018 4th international conference on a computing communication control and automation (ICCUBEA).

[2] Pingale, Kedar, et al. "Disease Prediction using Machine Learning." (2019).Mr. ChalaBeyene, Prof. Pooja Kamat, “Survey on Prediction and Analysis the Occurrence of Heart Disease Using Data Mining Techniques”, International Journal of Pure and Applied Mathematics, 2018.

[3] S. Patel and H. Patel, “Survey of data mining techniques used in healthcare domain,” Int. J. of Inform. Sci. and Tech., Vol. 6, pp. 53-60, March, 2016.

[4] Thirunavukkarasu K, Singh AS, Irfan M, Chowdhury A (2018) Prediction of liver disease using regression and classification algorithms. In: 2018 4th international conference on computing communication and automation (ICCCA) there will provide an automated predictor for diseases.

[5] Ray S (2019) A quick review of different machine learning algorithms. In: 2019 international conference on machine learning,ann, big data, cloud and parallel computing (Com-IT-Con), India, 14th–16th Feb 2019.