Disease Prediction Based on Symptoms Using Machine Learning

Y. Deepthi, K. Pavan Kalyan, Mukul Vyas, K. Radhika, D. Kishore Babu, and N. V. Krishna Rao

Abstract The healthcare domain is one of the prominent research fields in the current scenario with the rapid improvement of technology and data. It is difficult to handle the huge amount of data of the patients. It is easier to handle this data through Big Data Analytics. There are a lot of procedures for the treatment of multiple diseases across the world. Machine Learning is an emerging approach that helps in prediction, diagnosis of a disease. This paper depicts the prediction of disease based on symptoms using machine learning. Machine Learning algorithms such as Naive Bayes, Decision Tree and Random Forest are employed on the provided dataset and predict the disease. Its implementation is done through the python programming language. The research demonstrates the best algorithm based on their accuracy. The accuracy of an algorithm is determined by the performance on the given dataset.

Keywords Big data analytics · Machine learning algorithms · Decision tree · Random forest · Healthcare · Naive Bayes · Python

1 Introduction

As we know that health is the priority even before the technology exists. Talking about the research field in terms of healthcare, there is a lot of scopes and it is evolving at a faster rate. To upgrade the existing model or technology, research on healthcare will affect totally in terms of humanity. Nowadays, healthcare communities are storing patient data so that they can use in the research area and can be used in developing new technology. Storing the tremendous data involved in big data and where big data analytics comes into action [1].

Big data is consists of humongous with complex data which is difficult to understand. It provided by the organization to store, handle and manage the data properly. To understand big, it also provides the solution to store and analyze unstructured and structured data in an effective manner rather than using RDMS (relational database

Y. Deepthi (✉) · K. P. Kalyan · M. Vyas · K. Radhika · D. K. Babu · N. V. Krishna Rao
Department of CSE, Institute of Aeronautical Engineering, Dundigal, Hyderabad, Telangana, India
e-mail: deepthisagar7@gmail.com

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2020

A. Sikander et al. (eds.), *Energy Systems, Drives and Automations*, Lecture Notes in Electrical Engineering 664, https://doi.org/10.1007/978-981-15-5089-8_55

management system). Big data provides the solution to handle the problem occurs in medical in term of data and prediction analysis is always being better when it comes to big data [2]. Big data Analytics is applied often in healthcare and it is used to handle and analyze the huge data [3] long term. Greater the time it takes to analyze the invasion, the greater are the chances that the system becomes vulnerable [1]. Any software or hardware which has higher computation capacity it least time is chosen over other existing tools.

Machine learning is a program that is used to perform some tasks. The machine takes some decision and predicts results [4]. The machine learning topic is getting popular which is used to solve real-time examples. Coming to the healthcare topic, prediction of the disease can be done by using the machine learning algorithm [1].

Our paper is all about predicting diseases based on symptoms using three classifiers, i.e. random forest, decision tree and Naive Bayes. Here, while predicting the diseases we are finding the accuracy which will help to make a decision appropriately.

Our paper consists of a section which is as follows: Second section provides the analysis of the existed paper used for the proposed project. Third section is about the methodology for predicting diseases based on different algorithms. Fourth section gives the output of the proposed project. At last, the 5th section will provide a conclusion.

2 Literature Survey

This section provides a survey on the already existed system approaches and understands the challenges behind it. It also provides the loophole present in the existing system so that we can overcome those loopholes and make it as a proposed system. In this paper, it is based on the prediction of disease which is diabetes disease using machine learning algorithms. Four machine learning algorithms are used to predict diabetes disease on Weka tool. The four algorithms, i.e. Simple CART, Naive Bayes, SVM and random forest are used for prediction and analyze the diabetes data. All four algorithms are built to compare against each other by their accuracy. The best model for predicting the diabetes disease of a patient is SVM because SVM is giving the best result when we compare with other algorithms while predicting diabetes diseases [1].

In this paper, it is based on heart disease prediction using machine learning such as SVM, Naive Bayes and Decision Tree. All four classifiers are applied with and without using PCA (Principle component analysis) on the dataset. They are using PCA as it is used to reduce the attributes from the dataset. After reducing the dataset, they observed that SVM is performing better than Naive Bayes and random forest. So, SVM can be used to predict heart disease. The result of this paper is to predict the disease like heart disease and diabetes diseases using Weka tool [5].

In this manuscript, it's about the prediction of diabetes disease using data mining. According to this paper, KNN and Naive Bayes algorithms are being used for prediction of disease and comparing accuracy against each other based on the disease

dataset. For the prediction of disease, they were using the disease dataset and by analyzing and processing that dataset to produce the appropriate output. Datasets consist of around 2000 diabetic patient data. When they compare both algorithm Bayesian and KNN based on accuracy, both of them gives better accuracy on larger dataset than smaller datasets [6].

In this manuscript, it's about predicting liver disease using a classification algorithm, i.e. logistic regression, SVM and KNN have been used to predict liver disease. All algorithms are being compared based on their accuracy through a confusion matrix. The liver disease dataset being downloaded from UCI with an instance of 567, whereas data is collected from the ILPD (Indian Liver Patient dataset). From the collected result and after analyzing of the result, they got to know that two algorithms, i.e. KNN and Logistic regression is having the best accuracy compared with other algorithms and out of these two algorithms, logistic regression is best and highly responsive in term of true positive rate or recall. So, the best model for liver disease prediction is logistic regression which is giving the best accuracy among all other algorithms [7].

In this paper, the author was telling about the heart disease is being predicted based on the machine learning algorithm. This paper consists of two algorithms, i.e. decision tree and Naive Bayes. Now, they are comparing those algorithms based on their accuracy and they are using the python programming language which helps them to find the accuracy with the appropriate result and helps them to figure out which algorithm is best for their proposed model. Dataset is being downloaded from the UCI machine learning repository containing 300 instances. Therefore, the result is to check whether a person having heart disease or not based on two classifiers. These two algorithms are producing accuracy one of them is the decision tree with 91% and Naive Bayes with 87% [8].

3 Methodology

This topic describes the approaches that we are using to build our project and how we are analyzing our project.

3.1 Anaconda Python Software

The features of anaconda software are as follows:

It used in machine learning applications, data science, predictive analytics, etc.
Developed by anaconda, Inc. and written in python.
The following software's are provided in Anaconda Navigator are as follows:

JupyterLab
JupyterLab Notebook

R-Studio
Spyder, etc

The current version of anaconda software is used based on the python version, i.e. 3.7.

3.2 Disease Dataset

The disease dataset is used and downloaded from the Kaggle website. Approximately 4000 instances are present in our disease data set. Attributes describe symptoms and prognosis which means it contains diseases. Datasets are divided into two set testing sets and training set with a size of 13.5 KB and 1.4 MB, respectively. About the disease dataset, it is used to find disease based on symptoms and these data are being used by the machine learning algorithms so that prediction can be done with correct accuracy.

4 Flow Chart of Our Proposed System

The proposed model is made for disease prediction based on symptoms and it takes the input from the dataset for symptoms and its correspondence diseases. The data, i.e. the symptoms that are entered by the user is taken by the software are being processed by the algorithms and these algorithms will analyze the data and produce the accuracy and come to a conclusion and show the result, i.e. disease based on symptoms. In the proposed model, the result is produced after processing and analysis of algorithms and we will print the disease with accuracy given by each algorithm.

Figure 1 shows the method that has been applied by our proposed model.

5 Algorithms

This subtopic describes the algorithms that are used in our proposed system which as follows: Naive Bayes Algorithm, Decision Tree Algorithm and Random Forest Algorithm.

5.1 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. The method

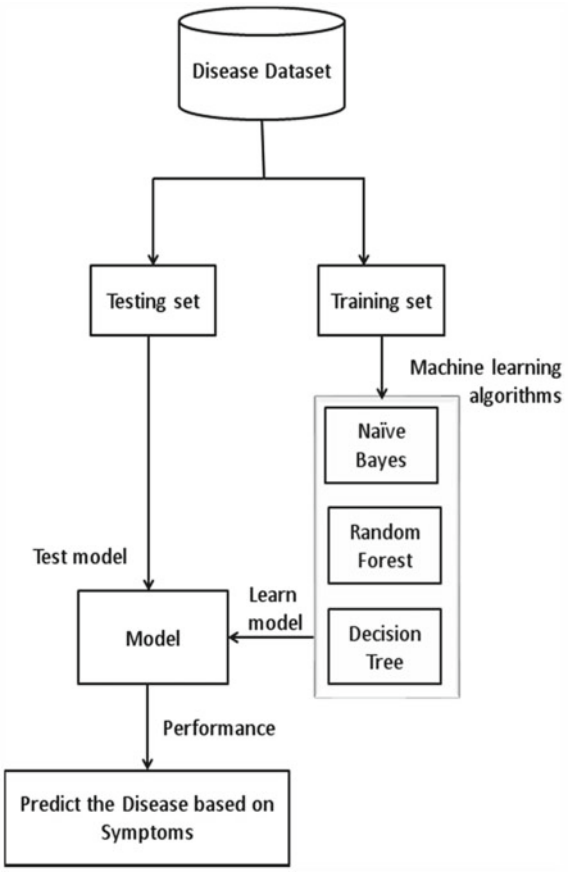

Fig. 1 Low representation of the proposed model

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

$$P(H|E) = (P(E|H) * P(H))/P(E) \tag{1}$$

$P(H)$ = 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

5.2 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.
2. 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.
3. 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 or not.

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 called

decision 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.

6 Result and Discussion

In our paper, we are checking what type of disease a person has based on symptoms through machine learning. Here, our project will predict the disease by applying machine learning algorithms namely decision tree, Naive Bayes and Random Forest. These algorithms predict a similar disease with different accuracy.

The accuracy of the random forest is 94.6% which is highest among all other algorithms. The accuracy of Naive Bayes is 84.5% which is highest than the decision tree and the accuracy of the decision tree is 78.5%. Overall, we can get to know that random forest is effective in terms of predicting diseases. Below bar graphs are used to compare different algorithms based on accuracy with different types of diseases (Figs. 2, 3 and 4).

From this experiment, we got to found out that with the highest accuracy the random forest is the best model compared to other algorithms. So, for the prediction of diseases, we shall use the random forest as a default model to give better results and accuracy.

Fig. 2 Comparison between three different algorithms while predicting Jaundice based on symptoms and from the graph, we got to know that random forest is having the highest accuracy

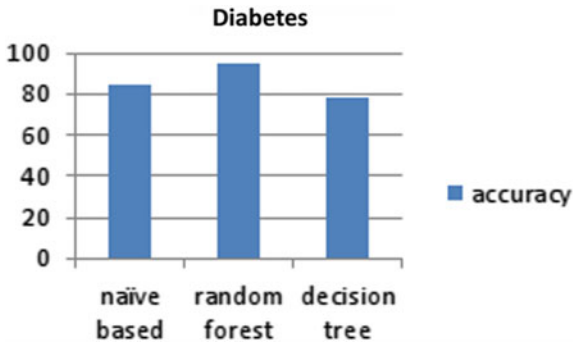

Fig. 3 Comparison between three different algorithms while predicting diabetes based on symptoms and from the graph, we got to know that random forest is having the highest accuracy

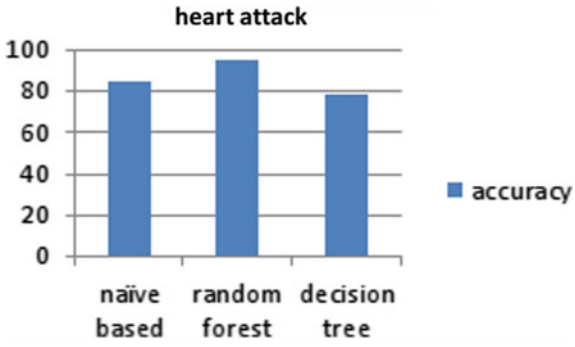

Fig. 4 Comparison between three different algorithms while predicting heart attack based on symptoms and from the graph, we got to know that random forest is having the highest accuracy

7 Conclusion

From our research paper, we conclude that we are using three machine learning algorithms, i.e. random forest, Naive Bayes and decision tree for predicting diseases based on symptoms. These algorithms are used to predict disease based on symptoms and are being compared through their accuracy against each other. From the experiment, we get to know that random forest has the highest accuracy among other algorithms.

Therefore, it can conclude that random forest is appropriate for predicting diseases based on symptoms.

In the future, we can upgrade our project by adding some other machine learning algorithms and can compare their accuracy against each other so that it will be easy to find out the disease based on symptoms.

References

1. Mir A, Dhage SN (2018) Diabetes disease prediction using machine learning on big data of healthcare. In: 2018 fourth international conference on computing communication control and automation (ICCUBEA)
2. Singh M, Bhatia V, Bhatia R, Big data analytics solution to healthcare
3. Koppad SH, Kumar A, Application of big data analytics in healthcare system to predict COPD
4. Ray S (2019) A quick review of machine learning algorithms. In: 2019 international conference on machine learning, big data, cloud and parallel computing (Com-IT-Con), India, 14th–16th Feb 2019
5. DhomseKanchan B, MahaleKishor M (2016) Study of machine learning algorithms for special disease prediction using principal component analysis. In: 2016 international conference on global trends in signal processing, information computing and communication. IEEE 20168
6. Shetty D, Rit K, Shaikh S, Patil N (2017) Diabetes disease prediction using data mining. In: 2017 international conference on innovations in information, embedded and communication systems (ICIIECS)
7. Thirunavukkarasu K, Singh AS, Irfan M, Chowdhury A (2018) Prediction of liver disease using classification algorithms. In: 2018 4th international conference on computing communication and automation (ICCCA)
8. SanthanaKrishnan J, Geetha S (2019) Prediction of heart disease using machine learning algorithms. In: 2019 1st international conference on innovations in information and communication technology (ICIICT)