Comparative Study of Different Machine Learning Algorithms For Prediction of Disease

Pranav Kumar, Rajesh Kumar, Md. Abid, Mayank Anand

G.L Bajaj Institute of Technology and Management

Greater Noida, India

Abstract

Machine Learning is one of the most important tool in the field of medical science for disease prediction. As any technology matures by time machine learning also achieved to its remarkable level of accuracy .With the help of different machine learning algorithms health Workers or the patient himself detect the disease and patient can get the proper treatment .It can not only detect the disease early but early detection can reduce the cost of treatment as well. In this study a comparative analysis on patients symptom datasets was performed to know the accuracy of different algorithms such naïve bays ,logistic regression support vector machine ,k-Nearest Neighbor ,decision tree and random forest. The accuracy for naïve bays is 85.25, Logistic regression is 85.25%, support vector machine is 81.97, k-Nearest Neighbor 67.21%, decision tree is 81.97 and random forest is 95.08%.

Keywords- Logistic regression, SVM, Random forest

I. Introduction

Machine Learning is growing day by day to achieve its precise accuracy as the size the data sets increase very rapidly. It is the subfield of artificial intelligence that allows machine to learn without explicit programming by exposing it to datasets by learning by its own experience. Data used for machine learning is of two types' labelled data and unlabeled data. Labelled data is a data where attributes are provided to some sort tag value and used for supervised learning. Unlabelled data is data where no labelling only data points are assigned to it. It is used for unsupervised learning so that machine can identify the pattern and structure present in the data.

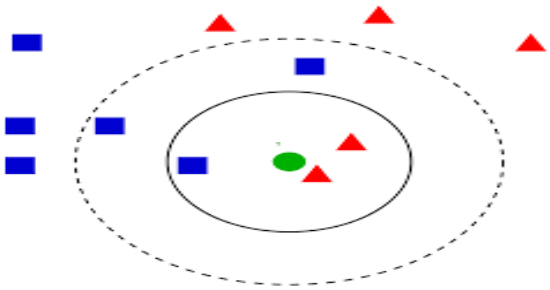
In this paper, six ML algorithms applied to a disease data set are compared on the basis of their accuracy. These algorithms are Naïve Bays, Logistic regression, support vector machine (SVN), k-Nearest Neighbor (KNN), Decision tree (DT) and Random Forest. The rest of the paper is organized as follows: Section II explains the fundamental concept of the six ML algorithms being compared on the basis of accuracy. Section III describes the accuracy based on the experimental observation. Finally, conclusions are provided in Section IV.

II. Machine learning Algorithms

For the purpose of comparative analysis, six Machine Learning algorithms are discussed. The different Machine Learning (ML) algorithms are k-Nearest Neighbor (KNN), Random Forest (RF), Support Vector Machine (SVM), Logistic regression, Decision tree, Naïve Bayes and . The reason to choose these algorithms is based on their popularity.

A. K-Nearest Neighbour(KNN)

Nearest Neighbour algorithms are among the simplest of all machine learning algorithms. The idea is to memorize the training set and then to predict the label of any new instance on the basis of the labels of its closest neighbours in the training set. The rationale behind such a method is based on the assumption that the features that are used to describe the domain points are relevant to their labelling in a way that makes close-by points likely to have the same label. Furthermore, in some situations, even when the training set is immense, finding a nearest neighbour can be done extremely fast.

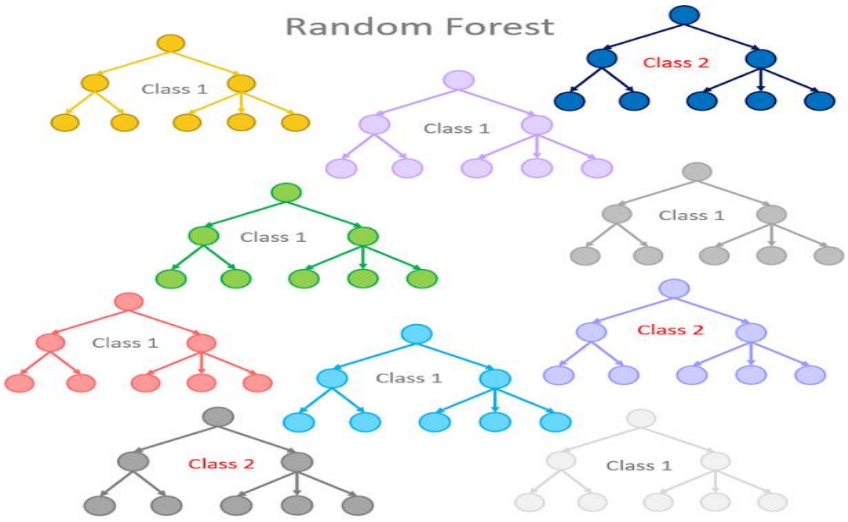


B. Random Forest

A random forest is a classifier consisting of a collection of decision trees, where each tree is constructed by applying an algorithm A on the training set S and an additional random vector, θ where θ is sampled i.i.d.(independently and identically distributed) from some distribution. The prediction of the random forest is obtained by a majority vote over the predictions of the individual trees.

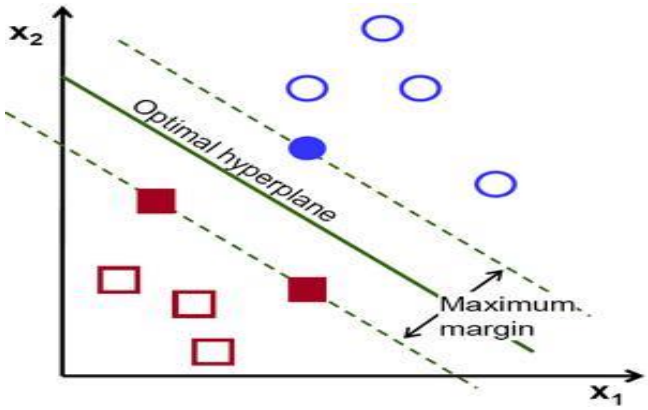
The Random Forest algorithm works in following steps:

1. Picks random K data points from the training data.
2. Builds a decision tree for these K data points.
3. Chooses the Ntree subset from the trees and performs step 1 and step 2
4. Decides the category or result on the basis of the majority of votes.



C. Support Vector Machine (SVM)

A Support Vector Machine (SVM) is a classifier which distinct the various classes of data by the use of a hyper-plane. SVM is modelled with the training data and it outputs the hyper-plane in the test data. The SVM model tries to find the space in the matrix of data where different classes of data can be widely differentiated and draws a hyper-plane.

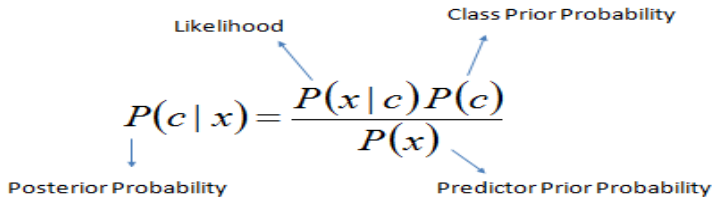


An optimal hyper-plane is chosen which maximizes the margin between the classes. Hyper-plane need not always be linear. A hyper-plane in SVM can also work as a non-linear classifier using technique known as kernel-trick.

D. Naïve Bayes

Naïve Bayes or Naïve Bayes classifier in a machine learning context is a classifier which uses the Bayes theorem to classify the data and it assumes that the probability of certain feature X is totally independent of another feature Y.

There are three types of Naïve Bayes. Gaussian Naïve Bayes, Multinomial Naïve Bayes and Bernoulli Naïve Bayes. Gaussian Naïve Bayes is used in classification problems, Multinomial Naïve Bayes is used in multinomial distributed data and Bernoulli Naïve Bayes is used in data with multivariate Bernoulli distribution.



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

E. Decision Tree

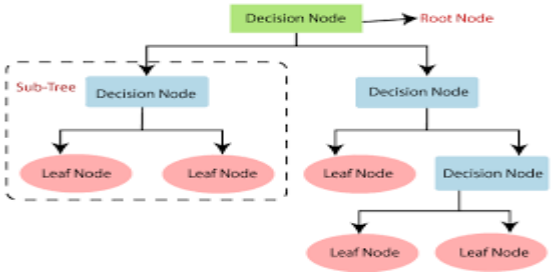
A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules.

In decision analysis, a decision tree and the closely related influence diagram are used as a visual and analytical decision support tool, where the expected values (or expected utility) of competing alternatives are calculated.

A decision tree consists of three types of nodes:[1]

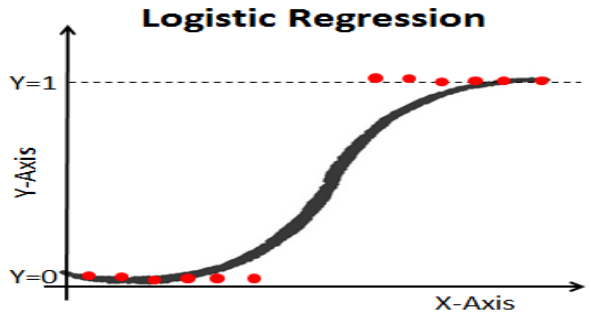
1. Decision nodes – typically represented by squares
2. Chance nodes – typically represented by circles
3. End nodes – typically represented by triangles

Decision trees are commonly used in operations research and operations management. If, in practice, decisions have to be taken online with no recall under incomplete knowledge, a decision tree should be paralleled by a probability model as a best choice model or online selection model algorithm. Another us of decision trees is as a descriptive means for calculating conditional probabilities.



F. Logistic Regression

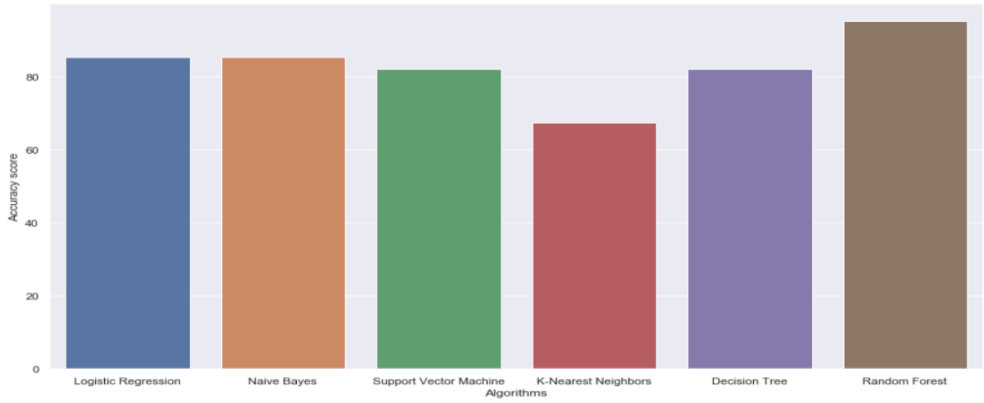
Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.



III. Comparative Analysis

Applying the above six proposed algorithms on our machine learning model with given datasets we get the accuracy

```
The accuracy score achieved using Logistic Regression is: 85.25 %
The accuracy score achieved using Naive Bayes is: 85.25 %
The accuracy score achieved using Support Vector Machine is: 81.97 %
The accuracy score achieved using K-Nearest Neighbors is: 67.21 %
The accuracy score achieved using Decision Tree is: 81.97 %
The accuracy score achieved using Random Forest is: 95.08 %
```



IV. Conclusion

ML techniques have been widely used in the medical field and have served as a useful diagnostic tool that helps physicians in analyzing the available data as well as designing medical expert systems. This paper presented six of the most popular ML techniques commonly used for disease detection and diagnosis, namely Support Vector Machine (SVM), Random Forest (RF) and Bayesian Networks (BN),Logistic regression ,Naïve Bayes ,K-Nearest Neighbour(KNN),Decision Tree(DT). The main features and methodology of each of the six ML algorithms was described. Performance comparison of the investigated techniques has been carried out using the Disease Data set. Experimental results obtained has proved that classification performance varies based

on the method that is selected. Results have showed that Random Forest have the highest performance in terms of accuracy, specificity and precision. However, RFs have the highest probability of correctly classifying disease.

REFERENCES

[1] WHO — Breast Cancer: Prevention and Control (2015) Retrieved 20 Jan 2015, from WHO — World Health Organization. http:// www.who.int/cancer/detection/breastcancer/en/index1.html

[2] Y. Elobaid, T.-C. Aw, J. N. W. Lim, S. Hamid, and M. Grivna, “Breast cancer presentation delays among Arab and national women in the UAE, a qualitative study,” SSM - Popul. Heal., Mar. 2016.

[3] E. D. Michie, D. J. Spiegelhalter, and C. C. Taylor, “Machine Learning , Neural and Statistical Classification,” Proceeding, 1994

[4] UCI Machine Learning Repository [online]. URL:https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/heart-disease.names

[5] Max Bramer, Principles of Data Mining. 2nd ed. Springer; 2013

[6] Robert A Wilson. The MIT Encyclopaedia of Cognitive Sciences. MIT Press; 1999

[7] Ethan Alapydin. Introduction to Machine Learning, 2nd ed. Cambridge Massa-chusetts, MIT Press:2010.

[8] Md Rahat Hossain, The Combined Effect of Applying Feature Selection and Pa-rameter Optimzation on Machine Learning Techniques for Solar Power Predic-tion. American Journal of Energy Research 2013; 1(1): 7-16.

[9]Gareth James. Introduction to Statistical Learning. New York, Springer; 2013.Ömer Cengiz Çelebi.

[10]. Principal Component Analysis [online]. 26 February 2002 URL:http://www.byclb.com/TR/Tutorials/neural_networks/ch5_1.htm Accessed 26 February 2017