IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A STUDY ON SUPERVISED MACHINE LEARNING ALGORITHMS WITH RESPECT TO PERFORMANCE ANALYSIS IN LIVER DISEASE DETECTION

[1] Dr. Netra Patil, [2]Ms. Rubina Sheikh

[1]Professor, [2]Assistant Professor
[1]Department of Master of Computer Applications
[1]Sinhgad Institute of Business Administration and Research, Pune, India

Abstract: Machine learning has a great potential in healthcare industry in various tasks ranging from diagnosis to decision making. There are several machine learning algorithms available which are suitable for various tasks but selecting the best one is a challenge. For selecting the best algorithm, we conducted a study on supervised machine learning algorithms for detecting liver disease in patients. In this paper, we have discussed various supervised machine learning algorithms and analyzed performance of those algorithms for liver disease detection for Indian liver patient records. To implement the algorithms, Indian liver patient records data set was used with 583 instances with 10 attributes as independent variables and one as dependent variable for the analysis. From this research study, the results show that SVM gives the best results in liver disease detection as compared to Logistic Regression, K-Nearest Neighbours, Naive Bayes (NB), Random Forest and Neural Network algorithm**.**

***Index Terms* - SVM, Logistic Regression, Classification, K-Nearest Neighbours (KNN), Naive Bayes (NB), Random Forest(RF), Neural Network, Accuracy Score, Mean AUC Score, Root Mean Square Error (RMSE).**

I. INTRODUCTION

In the world, patients with liver disease have been widely increasing because of excessive consumption of alcohol, harmful gases' inhalation, contaminated food consumption, drugs etc. This research study is an attempt to select the best algorithm to detect liver disease and thereby help to reduce burden on doctors.

Machine Learning is a branch of computer science which focuses on research in interdisciplinary area demonstrating a good blend of several technical and science branches like artificial intelligence, mathematics, statistics, medical science, engineering etc. The main motive behind machine learning research is to find fast and efficient learning algorithms that could be implemented in application useful across numerous industries for predictions based on data. Machine learning algorithms are mainly grouped into three categories – Supervised, Unsupervised and Reinforcement learning algorithms. Supervised machine learning algorithms make use of labeled data consisting of input value and a target output value as training data for analysis and derives an inferred function from it for mapping new target values. Unsupervised machine learning technique helps to identify hidden patterns from unlabelled data sets. Reinforcement learning algorithm makes a machine learn its behavior from the feedback received through the interactions with the external environment. Supervised and unsupervised learning techniques are generally used for data analysis and reinforcement techniques are used for decision making. This paper focuses on six supervised machine learning algorithms namely Logistic Regression, K-Nearest Neighbours, Naive Bayes (NB), Support Vector Machine (SVM), Random Forest and Neural Network used in medical diagnosis. In this paper, we have used Indian liver patient records data set consists of 583 patients with 10 clinical feature attributes to train different machine learning algorithms. In this research, we compared the accuracy of different supervised machine learning algorithms..

II. LITERATURE REVIEW

Machine learning is a concept in which machine learns from the past experience for improving future performance. In machine learning, algorithms are developed that uses data to learn themselves. Machine Learning can be effectively used in the medical field for the diagnosis and treatment by doctors. Medical diagnosis by screening at early stage is very important and beneficial as further treatment can be started earlier. With the clinical data available, the machine learning techniques can be used to aid in early diagnosis of the diseases like diabetes, cancer, heart, kidney, liver problem etc. In this section, we are discussing six supervised machine learning algorithms namely Logistic Regression, K-Nearest Neighbours, Naive Bayes (NB), Support Vector Machine (SVM) and Random Forest that are commonly Supervised Learning algorithms used for medical diagnosis.
Supervised Learning technique is a simpler technique compared to unsupervised learning yet highly accurate and trustworthy method. Supervised Learning algorithms make the machine learn using labeled data where results are already predicted and help to predict the results for unforeseen data. The result can be a continuous value or a discrete value. Mostly supervised learning algorithms are categorized as

Regression and Classification algorithms. Regression algorithm predicts a single output value using training data whereas Classification algorithm predicts the output in the form of category.

A. Logistic Regression

Logistic Regression is used for the classification and it is a predictive analysis algorithm based on the concept of probability. It finds the best fit parameter to estimate the probability of the binary response based on one or more features using the logistic sigmoid function [12].

B. Support Vector Machine (SVM)

SVM is a supervised learning algorithm mainly used for classification where the data is linearly classified by constructing hyper planes in a multidimensional space that distinctly classifies the data points. This plane is called as optimal hyper plane. The objective of the algorithm is to find the optimal plane with the maximum margin, i.e the maximum distance between data points of both classes. Hyperplanes are the decision boundaries which help classify the data points. The dimension of the hyperplane depends upon the number of features. The data points near the hyperplane are called as support vectors. These support vectors can be used to increase the margin between classifiers. Hyperplane position can be changed by deleting these support vectors.[19].

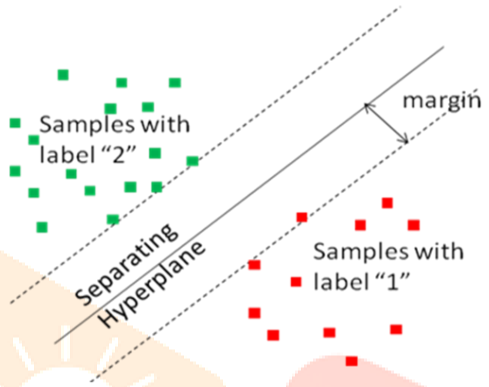

Figure 1 Maximum margin, the vectors on the dashed line are the support vectors

C. Naive Bayes (NB)

Naïve Bayes is a supervised learning method based on Bayes' Theorem that selects the decision based on conditional probability. The best thing about Naive Bayes is that it only requires a small number of training data to estimate the parameters necessary for classification. Naive Bayesian equation is used to calculate the posterior probability for each class. The outcome of prediction is the class with the highest posterior probability. It is called Naïve because the classifier assumes the features are independent.[1]

D. K-Nearest Neighbours

The K-Nearest Neighbour is a simple, easy to implement instance based supervised learning technique that classifies based on a similarity measure, like Euclidean, Mahanttan or Minkowski distance functions. The output can be calculated as the class with the highest frequency from the K-most similar instances. Each instance votes for their class and the class with the highest number of votes is taken as the prediction.[4]

E. Random Forest

Random forest is an ensemble learning technique for classification and regression which operates by constructing a large number of decision trees at the time of training and generates output as the class, that is the mode of the classes (classification) or mean prediction (regression) of the individual trees [16]. Random forest algorithm was proposed by Leo Breiman in 2001[3] which is also known as random decision forest. In the random forest algorithm, there are two steps, first is to create a random forest and then predict from the random forest classifier. In the random forest creation, k features are selected from the total of m features where k<m. Among the k features, a node i is selected by the best split point. Then, children nodes are split using the best split. This procedure is repeated until the specified count of nodes is reached. Repeating the steps similarly, we can build the forest at random subspace. After random forest creation, the test features are taken and the outcome for generated decision trees is predicted using the rules. The votes for each predicted target are calculated. The final prediction is the high voted predicted target. This algorithm is an enhancement to the decision tree algorithm. [5].

F. Neural Network

Neural Networks are a set of algorithms which are modeled on neurons in the brain and are designed to recognize patterns. NN consists of different layers for analyzing and learning data. These layers comprise of nodes, where computations occurs which in turn fires on receiving sufficient stimuli. A node combines input from the data with weight that either amplify or reduce that input, thereby assigning significance to inputs required for the task the algorithm is trying to learn. The artificial neurons are interconnected and communicate with each other. These input-weight products are summed and then the sum is passed through a node's activation function, to determine whether and to what extent that signal should progress further through the network to affect the ultimate outcome, i.e. Classification. If the signal passes through, the neuron has been activated. Each connection is weighted by previous learning events and with each new input of data more learning takes place. Each layer's output is simultaneously the subsequent layer's input, starting from an initial input layer receiving initial data. More the layers in a neural network, more it is learned and more accurate the pattern detection is.[8]

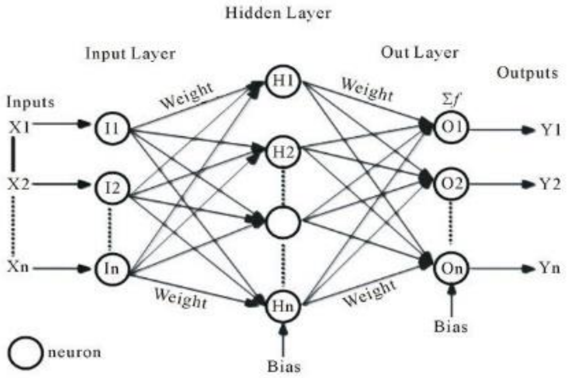

Figure 2 : Neural Network Layers

III. Proposed Work

Dataset Used

Indian liver patient records data set consists of 583 patients' records with 10 fields showing clinical features such as Age, Gender, Total_Bilirubin, Direct_Bilirubin, Alkaline_Phosphotase, Alamine_Aminotransferase, Aspartate_Aminotransferase, Total_Protiens, Albumin, Albumin_and_Globulin_Ratio. Dataset Label has two values 1 or 2 for classifying whether liver disease is present or not.

Table 1: Clinical features of each patient

Clinical feature	**Description**
Age	Age of the patients
Gender	Sex of the patients
Total_Bilirubin	Total Billirubin in mg/dL
Direct_Bilirubin	Conjugated Billirubin in mg/dL
Alkaline_Phosphotase	ALP in IU/L
Alamine_Aminotransferase	ALT in IU/L
Aspartate_Aminotransferase	AST in IU/L
Total_Protiens	Total Proteins g/dL
Albumin	Albumin in g/dL
Albumin_and_Globulin_Ratio	A/G ratio
Dataset	Label (patient has liver disease or not)

Experiments performed

Simulation experiments were carried out using open source Anaconda 3.0 distribution for implementing and evaluating various machine learning algorithms. We used sklearn preprocessing library for data preprocessing. We used Pandas to load the dataset and to perform calculations. For all the experiments train - test split size is 0.75 - 0.25.

Performance metrics to evaluate algorithms

To analyze the accuracy of the above algorithms, three performance metrics Accuracy Score, Mean AUC Score and Root Mean Square Error (RMSE) are used.

Accuracy Score – Accuracy is one of the metrics for evaluating classification algorithms. It is the ratio of number of correct predictions to the total number of input samples.

Mean AUC Score – AUC represents "Area under ROC curve" and it gives the rate of successful classification. ROC (Receiver Operating Characteristic) Curve shows how good the model can distinguish between two things.

Root Mean Square Error (RMSE) – RMSE is a standard way of measuring the error of the model in prediction of data. RMSE shows the deviation between predicted and observed values.

IV. Results and Discussion

The performances of six supervised machine learning algorithms, Logistic Regression, K-Nearest Neighbours, Naive Bayes (NB), Support Vector Machine (SVM), Random Forest and Neural Network against parameters as Accuracy Score, Mean AUC score and Root Mean Square Error are discussed below in Table 2:

Table 2: Performance Analysis of algorithms implemented

Algorithm used	**Accuracy Score**	**Mean AUC Score**	**Root Mean Square Error**
Logistic Regression (LR)	0.726027	**0.730462**	0.523424
Naïve Bays (NB)	0.630137	0.699032	0.608164
K-Nearest Neighbours (KNN)	0.684931	0.653138	0.561310
Support Vector Machine (SVM)	**0.739726**	0.632529	**0.510171**
Random Forest (RF)	0.705480	0.701280	0.542698
Neural Network (NN)	0.719178	0.695812	0.529926

From Table 2, it is observed that Support Vector Machine (SVM) gives the best result on the basis of Accuracy score while predicting liver disease in patients as compared to rest of the algorithms whereas Logistic Regression (LR) gives best result on the basis of Mean AUC score. Also it can be seen that Root Mean Square Error (RMSE) value of SVM is the lowest amongst all algorithms which interprets that the performance of SVM is the best.

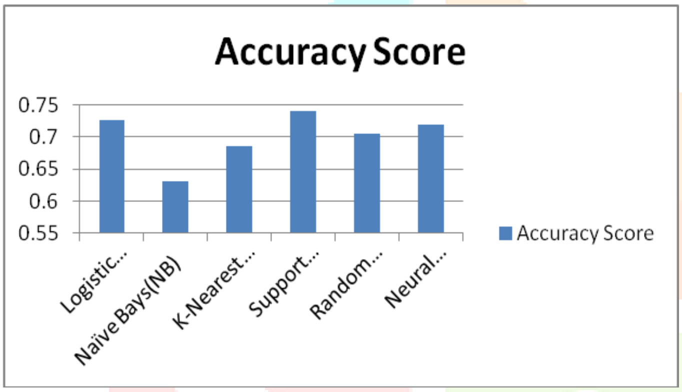

Figure 1 Accuracy Scores of implemented algorithms

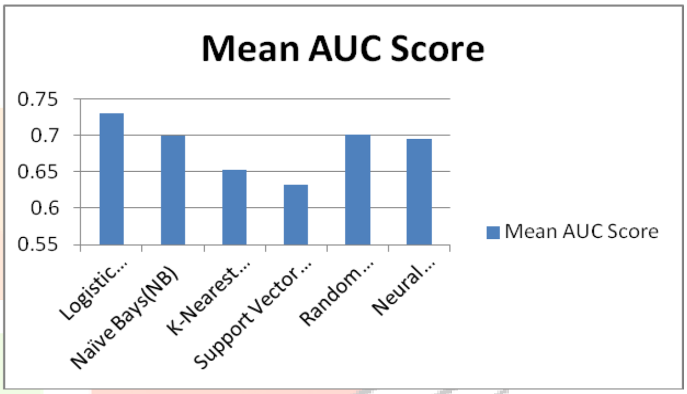

Figure 2 Mean AUC Scores of implemented algorithms

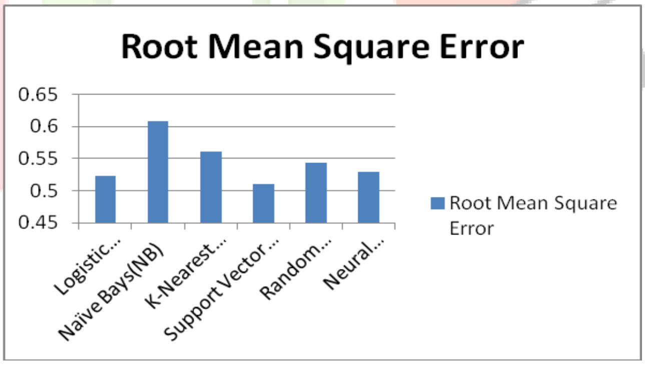

Figure 3 RMSE values of implemented algorithms

V. CONCLUSION

The main purpose of this paper is to check whether the liver disease is present in patients or not at an early stage based on their clinical data using supervised machine learning algorithms. The performance of various supervised machine learning algorithms was compared on the basis of accuracy score, mean AUC score, RMSE value and finally it is concluded that Support Vector Machine (SVM) algorithm gives better accuracy. In future, performance improvement of these algorithms can also be checked after parameter tuning.

VI. REFERENCES

[1] A. Krishna, D. Edwin and S. Hariharan, "Classification of liver tumor using SFTA based Naïve Bayes classifier and support vector machine," 2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT), Kannur, 2017, pp. 1066-1070.

[2] Ajinkya Kunjir, Basil Shaikh, “A Survey on Machine Learning Algorithms for Building Smart Systems” in International Journal of Innovative Research in Computer and Communication Engineering, 5, 1052- 1058, January 2017.

[3] Breiman, L. Machine Learning (2001) 45: 5.

[4] Chich-Min Ma, Wei-Shui Yang and Bor-wen Cheng, "How the parameter of k-nearest neighbours Algorithm impact on the Best Classification Accuracy: In case of Parkinson Dataset", Journal of applied sciences, Vol. 14 (2), pp.171-176,2014.

[5] Cortes C, Vapnik V (1995) Support-vector networks. Machine Learning 20(3):273–297

[6] D. H. Abd and I. S. Al-Mejibli, "Monitoring System for Sickle Cell Disease Patients by Using Supervised Machine Learning," 2017 Second Al-Sadiq International Conference on Multidisciplinary in IT and Communication Science and Applications (AIC-MITCSA), Baghdad, Iraq, 2017, pp. 119-124.

[7] Gerard Biau, “Analysis of a Random Forests Model” in Journal of Machine Learning Research, 13, 1063-1095, 2012.

[8] https://chatbotsmagazine.com/

[9] Kotsiantis, S. B. (2007). Supervised Machine Learning: A Review of Classification Techniques. Informatica 31 (2007). Pp. 249 – 268. Retrieved from IJS website: http://wen.ijs.si/ojs-2.4.3/index.php/informatica/article/download/148/140.

[10] L. Liu, "Research on Logistic Regression Algorithm of Breast Cancer Diagnose Data by Machine Learning," 2018 International Conference on Robots & Intelligent System (ICRIS), Changsha, 2018, pp. 157-160.

[11] Liaw A, Wiener M, Classification and Regression by Random Forest, R News, Vol 2/3, Dec 2002

[12] M. Bozorgi, K. Taghva and A. Singh, "Cancer survivability with logistic regression," 2017 Computing Conference, London, 2017, pp. 416-420.

[13] M. F. Akay, "Support Vector Machines combined with feature selection for breast cancer diagnosis", Expert systems with applications, vol. 36, no. 2, pp. 3240-3247, 2009.

[14] Michael A. Nielsen, "Neural Networks and Deep Learning", Determination Press, 2015

[15] Murat KORKMAZ, Selami G NEY, ule Y ksel Y TER, "The importance of logistic regression implementations in the turkish livestock sector and logistic regression implementations/fields", J.Agric. Fac. HR.U., 2012.

[16] N. Peter, “Enhancing random forest implementation in WEKA”, in: Machine Learning Conference, 2005.

[17] S. Kaur and S. Kalra, "Disease prediction using hybrid K-means and support vector machine," 2016 1st India International Conference on Information Processing (IICIP), Delhi, 2016, pp. 1-6.

[18] S. Marshland, Machine Learning an Algorithmic Perspective. CRC Press, New Zealand, 6-7, 2009.

[19] Types of Machine Learning Algorithms, Taiwo Oladipupo Ayodele, University of Portsmouth, United Kingdom.