Performance Evaluation of Various Classification Algorithms to Predict the Liver Disease of a Patient.

K.S.S.M Ravi Kiran[1], B.Vikas[2]
[1]*UG Student,* [2]*Assistant Professor*
Computer Science and Engineering, GITAM Institute of Technology, GITAM
Visakhapatnam, India

Abstract **-** Factors like environment quality, food habits, and alcohol consumption habits have led to a steep rise in patients with liver diseases. A substantial growth in the number of patients would mean an increased dependency on automated machines. Using these, doctors can easily judge the conditions of the patient. This paper assesses the efficiency of data mining classification algorithms to efficiently classify the dataset. The efficiency of all the algorithms we used are weighed based on accuracy, precision, sensitivity and specificity. We have trained the algorithms using the dataset by considering various attributes that affect the patient suffering from a liver disease. After training the classification algorithms we can effective predict the results.

Keywords **-** Data mining, Classification algorithms, Liver diagnosis.

I. INTRODUCTION

Data mining in medical diagnosis has been giving its best in aiding the patients for prevention and curing of their diseases based on their medical and genetic records[1]. With this much abundance of data present which has been extensively collected from various sources . The target must be getting the full functionality of the data and transform it into information that can be used efficiently[2].

Liver diseases can be caused by viruses, alcohol consumption and various environmental factors. As time progresses damage to the liver may lead to liver failure which can be considered as a life threatening condition.

It's quite difficult to discover the problem in the initial stages of the disease. To determine the health of the patient the enzyme levels in the blood are examined. Now these results are integrated with one of the classification techniques using the modern day technology to diagnose the patient in the initial stages.

The Indian Liver Patients dataset[3] we considered was initially pre-processed i.e., the data cleaning is performed on the data so that there are no discrepancies in the data. Later the outliers and repeating values are removed and then Naïve Bayes ,K-NN ,Decision tree ,Random forests ,ID3 ,Deep Learning ,MLP, Neural networks ,Support Vector Machines and Liner and Logistic Regression algorithms are performed on the data of the dataset which has a class label to predict whether the patient has the disease or not. Now the measures like accuracy are analyzed to get the efficiency of classification of the algorithm. Based on the accuracy of classification the best classification algorithm is selected based on the most productive results.

II. RELATED WORK

Data mining in Medical diagnosis can achieve very intriguing results. So these are some of the related works done in the field of medical diagnosis which include Analysis of Data Mining Techniques For Healthcare Decision Support System Using Liver Disorder Dataset by Tapas Ranjan Baitharu , Subhendu Kumar Pani [2] used decision trees J48, Naive Bayes, ANN, ZeroR, 1BK and VFI algorithm to classify these diseases and compare the effectiveness, correction rate among them.

Detecting Diseases in Medical Prescriptions Using Data Mining Tools and Combining Techniques by Teimouri, M., Farzadfar, F.SoudiAlamdari, M.Hashemi-Meshkini, A., Adibi Alamdari, P., Rezaei-Darzi, E., … Zeynalabedini [4] clearly explained to using the data mining tools with the help of prescriptions to identify the diseases.

Data mining in healthcare – a review by Jothi, N., Wahidah, H[5] has reviewed various methods that can be used in the process of knowledge discovery in the field of healthcare.

Application of the Apriori Algorithm for Prediction of Polycystic Ovarian Syndrome(PCOS) by Vikas B, B.S.Anuhya, K. Santhosh Bhargav, Sipra Sarangi[8], have found the strong relationships between the attributes using the association mining technique called Aproiri algorithm .

III. METHODOLOGY

We initially pre-processed the data by applying attribute selection methods.

As the first step of classification is learning as mentioned earlier. Various models constructed to classify the dataset which was taken from UCI machine learning repository. There are 11 attributes and 584 instances in the dataset. The data consists of attributes like Age, Gender, Total Bilirubin, Direct Bilirubin, Sgpt, Sgot, Albumin, Alkaline Phosphatase, Total Proteins, Ratio Albumin and Globulin Ratio and class label. We use the Naïve Bayes ,K-NN ,Decision tree ,Random forests ,ID3 ,Deep Learning ,MLP, Neural networks ,Support Vector Machines and Liner and Logistic Regression models for the data set.

Now we use the model to get different performance evaluation metrics for different classification algorithms and classify whether the patient has a liver disease or not.

All the pre-processing, classification and predictions are made using the Rapid Miner tool [6].

IV. DATA DESCRIPTION

In this section we describe the attributes used in our data set and what each attribute stands for and the description of each attribute.

Table 1: Data description of the attributes

Attribute names	Description
1. Age	Age of the patient
2. Gender	Gender of the patient
3. TB	Total Bilirubin
4. DB	Direct Bilirubin
5. Alkphos	Alkaline Phosphotase
6. Sgpt	Alamine Aminotransferase
7. Sgot	Aspartate Aminotransferase
8. TP	Total Protiens
9. ALB	Albumin
10. A/G	Ratio Albumin and Globulin Ratio
11. Selector field	Used to split the data into two sets (labeled by the experts)

V. CLASSIFICATION ALGORITHMS

Classification is a two-step process. Initially step can be considered as learning step in which we construct a classification model. Finally we use the model in the second step to predict the class labels for the given data[7].

Decision Tree Induction: A decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node holds a class label. The topmost node in a tree is the root node.

ID3: ID3 was developed by J. Ross Quinlan in the early 1970's which stands for Iterative Dichotomiser used to generate decision tree.

Naïve Bayes Algorithm: Naïve Bayes algorithm is mainly based on the Bayes theorem. It's one of the algorithms which can produce very high classification accuracy even on large datasets. This algorithm works on the assumption of c l ass conditional independence. Because of this it also exhibits the ability to learn quickly.

Random Forests: It is a supervised ensemble learning algorithm. Each of the classifier in it is a decision tree. Here the collection of classifiers is considered a forest and it generates individual decision trees by selecting the attributes at random at each node to split the tree.

Artificial Neural networks: The idea of ANN is based on the working of human brain. As we know that human brain consists of a number of neurons .Based on this a ANN is constructed with the input and output layers and hidden layers. Each transition from one layer to an another is assigned with a weight. Input is provided to the input layer and the output of the output layer is compared with the input which in return provides the error rate of the weights in the ANN.

Support Vector Machines: SVM is a supervised learning technique. The goal of SVM is to optimally separate the two classes using a determined hyper plane. We only use a single hyper plane to create a maximum margin between the two classes. SVM has a very training speed when compared to other classification algorithms.

Linear Regression: Linear Regression is the process of fitting the best line to the attributes present. Hence every attribute present in the dataset can be used to predict the other attributes.

K-NN: K Nearest Neighbour algorithm can be used for both classification and regression. In K-NN 'k' stands for the data set items that we consider for the classification process. We then calculate the desired distance which can be either Euclidian distance or Manhattan distance or any distance as per the users necessity.

Deep learning: Deep learning is sub field of machine learning which implements data mining using the artificial neural networks. Deep learning works by building architectures such as deep neural networks and deep belief networks which can used in fields like medical diagnosis, drug design and bioinformatics.

VI. PERFORMANCE MEASURES

The performance of the above classification techniques can be evaluated by the following metrics:

Sensitivity: It is the fraction of positive tuples that are correctly classified.

$$sensitivity = \frac{frequencyoftruepositives}{(frequencyoftruepositives + frequencyoftruenegatives)}$$

Specificity: It is the fraction of negative tuples that are correctly classified.

$$specificity = \frac{frequencyoftruenegatives}{(frequencyoftruepositives + frequencyoftruenegatives)}$$

Precision: The measure of true positives in contrary to the all positive results present.

$$sensitivity = \frac{frequencyoftruepositives}{(frequencyoftruepositives + frequencyoffalsepositives)}$$

Accuracy: The percentage or proportion of tuples that are correctly classified by the classifier.

$$sensitivity = \frac{frequencyoftruepositives + frequencyoftruenegatives}{(frequencyoftruepositives + truenegatives + falsenegatives + falsepositives)}$$

VII. RESULTS AND DISCUSSION

All the classification algorithms obtain the following performance metrics:

Table 2: Performance evaluation of the classifiers

Classifier	Accuracy	Precision	Sensitivity	Specificity
Decision tree	72.90	100.00	100.00	5.39
K-NN	100.00	100.00	100.00	100.00
Naïve Bayes	72.04	65.63	65.63	88.02
Random Forest	73.07	100.00	100.00	5.99
Gradient boosted				

trees	86.45	88.46	88.46	81.44
ID3	71.36	100.00	100.00	0.00
Deep Learning	69.64	65.87	65.87	79.04
MLP	71.36	85.82	85.82	35.33
Nueral Network	71.36	99.76	99.76	0.60
Linear regression	72.04	98.80	98.80	5.39
Logistic regression	100.00	100.00	100.00	0.00
SVM	71.36	100.00	100.00	0.00

The table(table 2) depicts the performance metrics of the classification algorithms Naïve Bayes ,K-NN ,Decision tree ,Random forests ,ID3 ,Deep Learning ,MLP, Neural networks ,Support Vector Machines and Liner and Logistic Regression models on the Indian Liver Patients dataset.

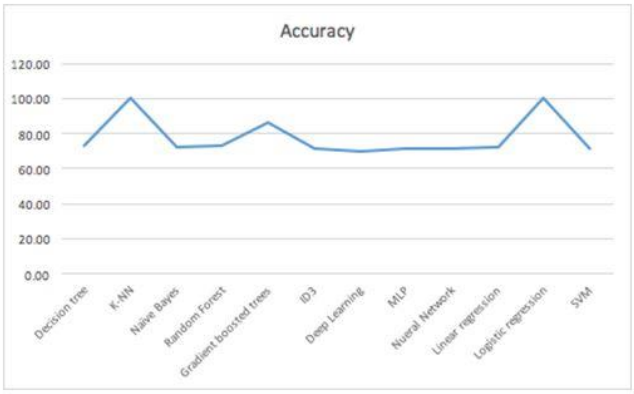

Figure 1: Graphical Representation of Accuracy

The graph (Figure 1) depicts that the K-NN and Logistic Regression classifiers have the highest accuracy when compared to the other algorithms. The algorithms showed a accuracy of about 100% mentioned classifiers were so efficient in predicting the health status of the liver of the patients.

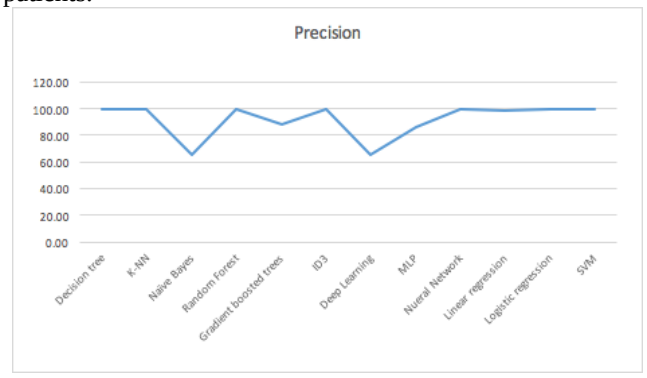

Figure 2: Graphical representation of Precision

The graph(Figure 2) depicts that the K- NN ,Decision tree ,Random forests ,ID3 ,Neural networks ,Support Vector Machines and Liner and Logistic Regression classifiers have the highest precision.

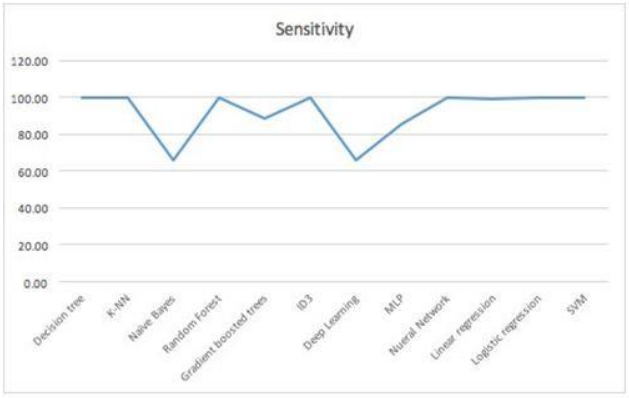

Figure 3: Graphical representation of Sensitivity

The graph(Figure 3) depicts that the K-NN ,Decision tree ,Random forests ,ID3 ,Neural networks ,Support Vector Machines and Liner and Logistic Regression classifiers have the highest sensitivity.

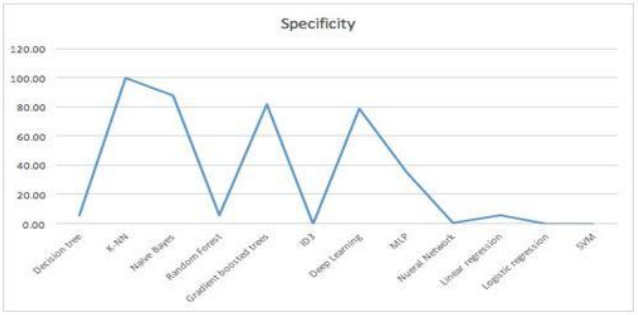

Figure 4: Graphical representation of Specificity

The graph(Figure 4) depicts that the K-NN has the highest specificity when compared to other classifiers.

VIII. CONCLUSION AND FUTURE SCOPE

With growing population of the world there is a huge need for efficient medical diagnosis and with the emerging technology there is huge scope for automated machines in health informatics[9].With the use of Naïve Bayes ,K-NN ,Decision tree ,Random forests ,ID3 ,Deep Learning ,MLP, Neural networks ,Support Vector Machines and Liner and Logistic Regression classifiers we have correctly classified with an accuracy of about 100%.This paper could be effectively implemented in hospitals for the diagnosis of patients with liver diseases. With initial discovery of the symptoms of the Liver diseases and treat the patients before its too late and prevent more serious and critical situations.
This paper has only implemented the classification algorithms on Indian Liver patient dataset. Therefore its scope can be increased and can be applied to more vast and

huge datasets on Liver diseases in the world and make predictions with good accuracies.

IX. REFERENCES

The template will number citations consecutively within

[1]. Uniqueness of medical data mining. Krzysztof J. Cios, G. William Moore Artif Intell Med. 2002 Sep-Oct; 26(1-2): 1–24.

[2]. Tapas Ranjan Baitharu ,Subhendu Kumar Pani (2016) Analysis of Data Mining Techniques for Healthcare Decision Support System Using Liver Disorder Dataset , https://doi.org/10.1016/j.procs.2016.05.27.

[3]. Indian Liver Patient dataset. UCI repository of machine learning databases. Available from https:// archive.ics.uci.edu/ml/machine-learning-databases/ 00225/, Last modified: 21-May-2012.

[4]. Teimouri M, Farzadfar F, Soudi Alamdari M, et al. Detecting Diseases in Medical Prescriptions Using Data Mining Tools and Combining Techniques. Iranian Journal of Pharmaceutical Research: IJPR . 2016;15(Suppl):113-123.

[5]. Jothi, N., Wahidah, H.: Data mining in healthcare – a review. Proc. Comput. Sci. 72, 306–313 (2015).

[6]. RapidMiner Studio8.2, Interactive Design. Products: RapidMiner, https://my.rapidminer.com/ .

[7]. Jaiven Han, Micheline Kamer, Jain pei ,3rd Edition, Data Mining Concepts and Techniques.

[8]. Vikas B., Anuhya B.S., Bhargav K.S., Sarangi S., Chilla M. Application of the Apriori Algorithm for Prediction of Polycystic Ovarian Syndrome (PCOS). In: Bhateja V., Nguyen B., Nguyen N., Satapathy S., Le DN. (eds) Information Systems Design and Intelligent Applications. Advances in Intelligent Systems and Computing, vol 672. Springer, Singapore(2018).

[9]. Wang J.T.L., Zaki M.J., Toivonen H.T.T., Shasha D. (2005) Introduction to Data Mining in Bioinformatics. In: Wu X., Jain L., Wang J.T., Zaki M.J., Toivonen H.T., Shasha D. (eds) Data Mining in Bioinformatics. Advanced Information and Knowledge Processing. Springer, London.

ABOUT THE AUTHORS:

Mr. K S S M Ravi Kiran is currently pursuing his Bachelor's degree from Department of CSE, GITAM since 2015. His research interests include Data Mining,Machine Learning and Neural Networks.

Mr Vikas B pursued Bachelor of Technology and Master of technology from JNTUH, Hyderabad. He is currently pursuing Ph.D. in the Department of CSE, GITAM. His main research work focuses on Deep Learning, Cryptography Algorithms, Machine Learning and Data Mining. He has years of teaching experience and 2 years of research experience.