IDENTIFICATION OF LIVER PATIENTS USING SUPERVISED LEARNING: A COMPARATIVE ANALYSIS

Dr. K. Vanaja
Assistant Professor
Department of Computer Science,
Adhiyaman College of Arts and Science for Women, Uthangarai, Krishnagiri Dt, Tamil Nadu, India

Abstract : Liver Failure is a serious condition and it affects the patient's life time. Disease identification is the most crucial task for treating any disease. Liver disease can be inherited genetically or caused by a variety of subjects that damage the liver. Machine learning technique is broadly used in various grounds of science and technology. They have been giving out meaningful information. It also explores in creation and study of algorithms which can learn from data. Data mining in healthcare is an evolving field of high importance for providing diagnosis and a deeper understanding of medical data. To build an effective disease management strategy, large amount of data should be analyzed for the early detection of the disease, assessment of the severity and early prediction of adverse events. This will impede the progression of the disease, will improve the quality of life of the patients and will reduce the associated medical costs. The aim of this research paper is to present the state-of-the-art of the machine learning methodologies applied for the prediction of liver failure. The main objective of this research work is to find the best classification algorithm in terms of precision, accuracy, specificity and sensitivity. Therefore, the present investigation was done to determine the relative performance of four classification algorithms namely, Support Vector machine (SVM), Logistic Regression, Random Forest and Decision Tree algorithm based on the available downloaded Indian Liver Patient Dataset(ILPD)**.**

IndexTerms **- : Liver, Machine Learning, Logistic regression, Random Forest, Decision Tree, Support Vector Machine.**

I. INTRODUCTION

The liver is the largest solid organ and the largest gland in the human body, that sits on the right side of the belly. Weighing about 3 pounds, the liver is reddish-brown in colour and feels rubbery to the touch. The liver has two large sections, called the right and the left lobes. The gallbladder sits under the liver, along with parts of the pancreas and intestines. The liver and these organs work together to digest, absorb, and process food [1].

Health care and medicine handles huge data on daily basis. **Liver failure** means that your **liver** is losing or has lost all of its function. It is a **life-threatening** condition that demands urgent medical care [2].

Liver disease is also referred to as hepatic disease. Liver disease is a large term that covers all the potential problems that cause the liver to fail to perform its designated functions. Usually, more than 75% or three quarters of liver tissue needs to be affected before a decrease in function occurs [3].

The liver's main job is to filter the blood coming from the digestive tract, before passing it to the rest of the body. The liver also detoxifies chemicals and metabolizes drugs. As it does so, the liver secretes bile that ends up back in the intestines. The liver also makes proteins important for blood clotting and other functions [4]. Figure 1 refers the structure of liver.

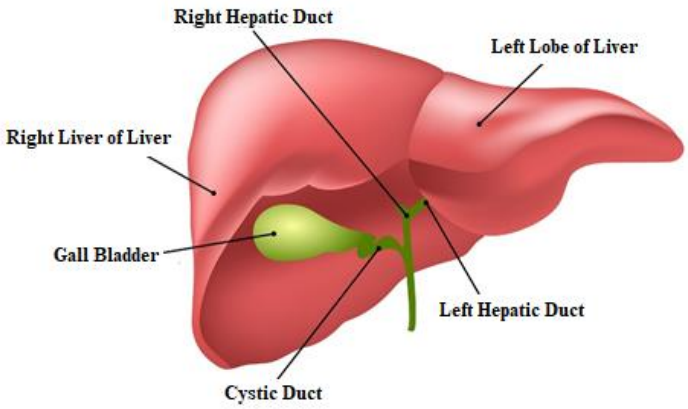

Figure 1. Structure of liver

1.1 Functions

The liver is classified as a gland and associated with many functions. It is difficult to give a precise number, as the organ is still being explored, but it is thought that the liver carries out 500 distinct roles [5]. Figure 2 shows the important functions of liver.

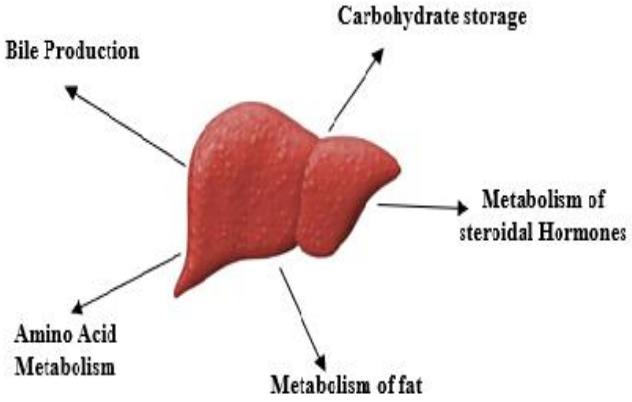

Figure 2. Functions of Liver

The major functions of the liver include:

- Bile production
- Absorbing and metabolizing bilirubin
- Supporting blood clots
- Fat metabolization
- Metabolizing carbohydrates
- Vitamin and mineral storage
- Filters the blood
- Immunological function
- Production of albumin
- Synthesis of angiotensinogen and
- Regeneration [6].

1.2 Liver Diseases

Liver disease is the occurrence of any trouble of liver function that causes sickness. The liver is responsible for most important functions of the body. If the liver fails to do those functions, it can cause significant injury to the body [7]. Liver disease is also referred as hepatic disease. The different types of liver diseases are largely classified according to the cause of the specific problem. some of which are acute and not serious while others are chronic and may be life-threatening [8].
The most common liver diseases [7] are:

- **Acute (sudden) hepatitis (inflammation)**: Acute hepatitis C is a contagious disease caused by the hepatitis C virus (HCV), which is spread through contact with infected blood and bodily fluids
- **Chronic (long duration) hepatitis**: This long-lasting liver infection is caused by the hepatitis C virus. It begins as an acute hepatitis that starts within the first 6 months of exposure to the virus.
- **Fatty liver disease**: steatosis, is a broad term that describes the build-up of fats in the liver. When too much fat builds up in your liver, that's fatty liver disease.
- **Cirrhosis (scarring)**: Cirrhosis is a late stage of scarring (fibrosis) of the liver caused by many forms of liver diseases and conditions, such as hepatitis and chronic alcoholism.
- **Cancer**: Cancers that affect the liver are most commonly metastatic cancers that have spread via the bloodstream to the liver from other sites in the body. However, primary cancers (cancers that arise in the liver) can also occur. The most common type of primary liver cancers is known as hepatocellular carcinomas.

The main objective of this research work is to classify the liver patients with the help of machine learning algorithms using the ILPD data set. Further this paper is organized with the following sections such as related work, machine learning techniques used, experimental evaluation and conclusion.

II **RELATED WORK**

Machine learning has attracted a huge amount of researches and has been applied in various fields in the world. In medicine, machine learning has proved its power in which it has been employed to solve many emergency problems such as cancer treatment, heart disease, dengue fever diagnosis and so on. Liver disease of the patients has been continuously increasing because of inhale of harmful gases, intake of contaminated food, different kinds of drugs and excessive consumption of alcohol. Automatic classification tools may reduce burden on doctors [9].

Bendi Venkata Ramana et al [10] evaluates the selected classification algorithms for the classification of some liver patient datasets. The classification algorithms considered here are Naïve Bayes classifier, C4.5, Back propagation Neural Network algorithm, and Support Vector Machines. These algorithms are evaluated based on four criteria: Accuracy, Precision, Sensitivity and Specificity.

Shapla Rani Ghosh and Sajjad Waheed et al [11] used the algorithms such as, Naive Bayes classification (NBC), Bagging, KStar, Logistic and REP tree were used to evaluate the accuracy, precision, sensitivity and specificity for diagnosing liver diseases. It was revealed that, KStar algorithm had the maximum accuracy, precision, sensitivity and specificity. On the

other, minimum accuracy was obtained from NBC. Therefore K* algorithm can be used on diagnosis tools or instruments for rapid identification of specific liver disorder.

Leoni Sharmila et al [12] provided a comparative study of different machine learning technique such Fuzzy logic, Fuzzy Neural Network and decision tree in classifying liver data set to predict the liver diseases. Fuzzy neural network results 91% accuracy.

Saranya et al [13] analysed the data of liver diseases using the classification techniques such as C4.5, Naive Bayes, Decision Tree, Support Vector Machine, Back Propagation Neural Network and Classification and Regression Tree Algorithms so as to find out the best classifier for manipulating the liver disorders. These algorithm gives various result based on speed, accuracy, performance and cost. It is seen that C4.5 gives better results compare to other algorithms.

Banupriya et al [14] implemented a feature model construction and comparative analysis for improving prediction accuracy of Indian liver patients. The outputs show from proposed classification implementations indicate that J48 algorithm performances all other classification algorithm with the help of feature selection with an accuracy of 95.04%.

Nazmun Nahar et al [15] calculated the performance of various decision tree techniques and compare their performance. The decision tree techniques used in this study are J48, LMT, Random Forest, Random tree, REPTree, Decision Stump, and Hoeffding Tree. The analysis proves that Decision Stump provides the highest accuracy than other techniques.

III MACHINE LEARNING ALGORITHMS

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. **Machine learning** focuses on the development of computer programs that can access data and use it learn for themselves [16]. The types of machine learning algorithms are as follows in figure 3.

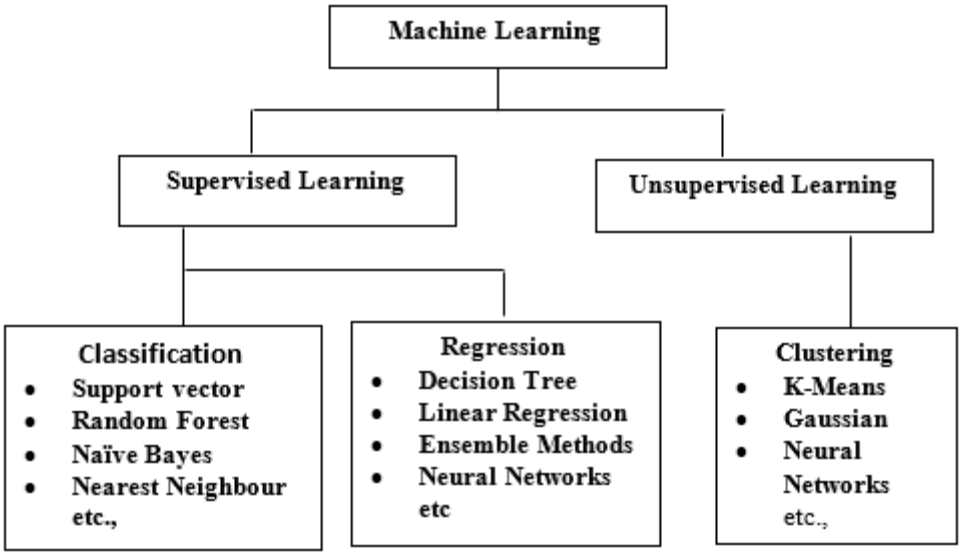

Figure 3. Types of Machine Learning Algorithms

3.1 Unsupervised:

It can be seen as a Machine Learning job used to draw derivations from datasets which contains input information without named responses. Cluster analysis is the most widely recognized unsupervised Learning technique. This technique is utilized for data examination to discover designs which are unseen [17].

3.2 Supervised:

It can be seen as a Machine Learning job of concluding a function from named training information. The training information will have an arrangement of preparing cases in which every instance is a combination of input object(typically a vector) and a required yield value(also called as supervisory flag).

This research work has been implemented and tested for comparison on supervised learning techniques[17] such as

- Logistic Regression
- Support Vector Machine
- Decision Tree
- Random Forest [17].

3.2.1 Logistic Regression

Logistic regression is another technique borrowed by machine learning from the field of statistics. It is the go-to method for binary classification problems (problems with two class values) [18].

Logistic Function

Logistic regression is named for the function used at the core of the method, the logistic function. The logistic function, also called the sigmoid function was developed by statisticians to describe properties of population growth in ecology, rising quickly and maxing out at the carrying capacity of the environment. It's an S shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits [18].

Formula

LR= 1 / (1 + e^-value)

Where e is the base of the natural logarithms and value is the actual numerical value that you want to transform. Logistic regression uses an equation as the representation, very much like linear regression. Input values (x) are combined linearly using weights or coefficient values (referred to as the Greek capital letter Beta) to predict an output value (y). A key difference from linear regression is that the output value being modelled is a binary value (0 or 1) rather than a numeric value [18].

3.2.2 Support Vector Machine

"Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. SVMs revolve around the notion of a "margin"- either side of a hyperplane that separates two data classes. Maximizing the margin and thereby creating the largest possible distance between the separating hyperplane and the instances on either side of it has been proven to reduce an upper bound on the expected generalization error [18].

In the SVM literature, a predictor variable which is called an attribute and a transformed attribute that is used to define the hyper plane is called a feature [18]. Here, choosing the most suitable representation can be taken as feature selection. A set of features that describes one case (i.e., a row of predictor values) is called a vector. The goal of this modelling is to find the optimal hyper plane which separates clusters of vector in such a way that cases with one category of the target variable are on one side of the plane and cases with the other category are on the other size of the plane. The vectors near the hyper plane are the support vectors.

3.2.3 Decision Tree

The decision tree is one of the most important and also most used classification algorithm. This algorithm utilizes a divide and conquer strategy to build a tree. There are a set of occurrences which are related with a collection of attributes [17]. A decision tree comprises of nodes and leaves in which nodes are tests on the estimations of a characteristics or attributes and leaves are the classes of an example that fulfils the given conditions. The outcome might be "true" or "false". Rules can be acquired from the way which begins from the root node and finishes at the leaf node and furthermore uses the nodes in transit as preconditions for the got rule, to foresee the class at the leaf. The tree pruning must be done to evacuate pointless preconditions and duplications [17]. There are two main types of Decision Trees:

1. **Classification trees** (Yes/No types)
   - Outcome was a variable like 'fit' or 'unfit'.
   - The decision variable is **Categorical**.

2. **Regression trees** (Continuous data types)
   - the decision or the outcome variable is **Continuous**, e.g. a number like 123.

3.2.4 Random Forest

Random Forest Classifier is ensemble algorithm. ***Ensembled algorithms*** are those which combines more than one algorithms of same or different kind for classifying objects. Random forest classifier creates a set of decision trees from randomly selected subset of training set. It then aggregates the votes from different decision trees to decide the final class of the test object [19].

All the above algorithms are used to classify the liver patients using ILPD dataset.

IV Data Set and its Details

Indian Liver Patient Dataset (ILPD) is considered to solve the problem of classification and comparison. This data set contains totally 583 Indian liver patients. In that 416 liver patient records and 167 non-liver patient records. This data set contains 441 male patient records and 142 female patient records. This dataset is downloaded from UCI Machine Learning Repository. The data set in .csv file format. The data set ILPD data set has the following attributes [20] tabulated in Table 1.

V EXPERIMENTAL EVALUATION

The performance all algorithms were analyzed with ILPD dataset. Among the classification algorithms Logistic Regression gave better accuracies with the considered attributes. Experimental evaluation is performed using the R Programming. The performance metrics considered for the evaluation is **Accuracy, Precision, Sensitivity and Specificity [21].**

The performance metrics [21] calculated by using four values, i.e., true positive, false positive, true negative and false negative are the basis to calculate all the measures. These values are described below.

TP = true positives: number of examples predicted positive that are actually positive
FP = false positives: number of examples predicted positive that are actually negative
TN = true negatives: number of examples predicted negative that are actually negative
FN = false negatives: number of examples predicted negative that are actually positive

Table 1. Attributes of ILPD with the selector field

Attribute	Description	Type
Age	Age of the patient	Integer
Gender	Gender of the patient	Categorical
TB	Total Bilirubin	Real Number
DB	Direct Bilirubin	Real Number
Alkphos	Alkaline Phosphotase	Real Number
Sgpt	Alamine Aminotransferase	Integer
Sgot	Aspartate Aminotransferase	Integer
TP	Total Protiens	Real Number
ALB	Albumi	Integer
A/G Ratio	Albumin and Globulin Ratio	Real Number
Selector field	used to split the data into two sets (labelled by machine learning algorithm)	

Accuracy: The accuracy of a classifier is the percentage of the test set tuples that are correctly classified by the classifier.

Accuracy = (TP+TN)/(TP+FP+TN+FN).

Precision: Precision is defined as the proportion of the true positives against all the positive results (both true positives and false positives)

Precision = TP/(TP+FP)

Sensitivity: Sensitivity is also referred as True positive rate i.e the proportion of positive tuples that are correctly identified.

Sensitivity = TP/(TP+FN)

Specificity: Specificity is the True negative rate that is the proportion of negative tuples that are correctly identified.

Specificity = TN/(TN+FP)

Table 2 gives the compared values of applied machine learning algorithms with the performance metrics such as accuracy, precision, sensitivity and specificity.

Table 2. Classification Algorithms with performance metrics

Classification Algorithms	Accuracy	Precision	Sensitivity	Specificity
Logistic Regression	81.9	0.840	0.954	0.625
SVM	80.7	0.166	0.523	0.411
Decision Tree	73.49	0.823	0.848	0.333
Random Forest	79.5	0.818	0.954	0.500

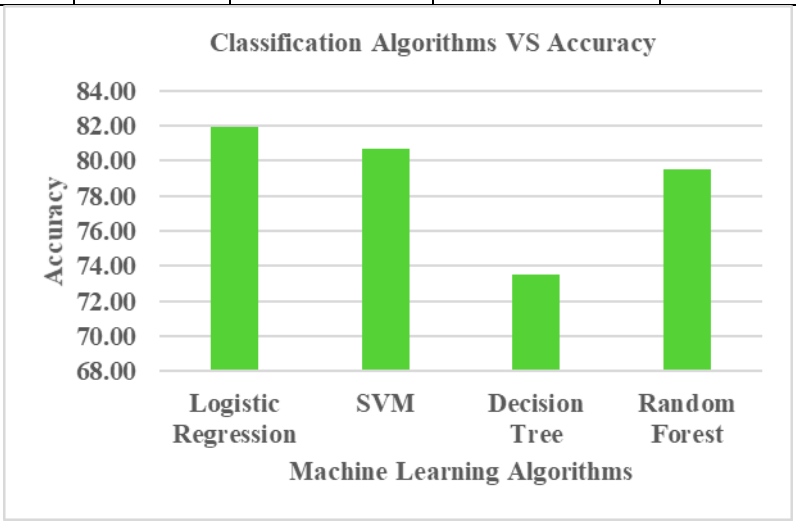

Figure 4. Comparison of Four Algorithms Vs Accuracy

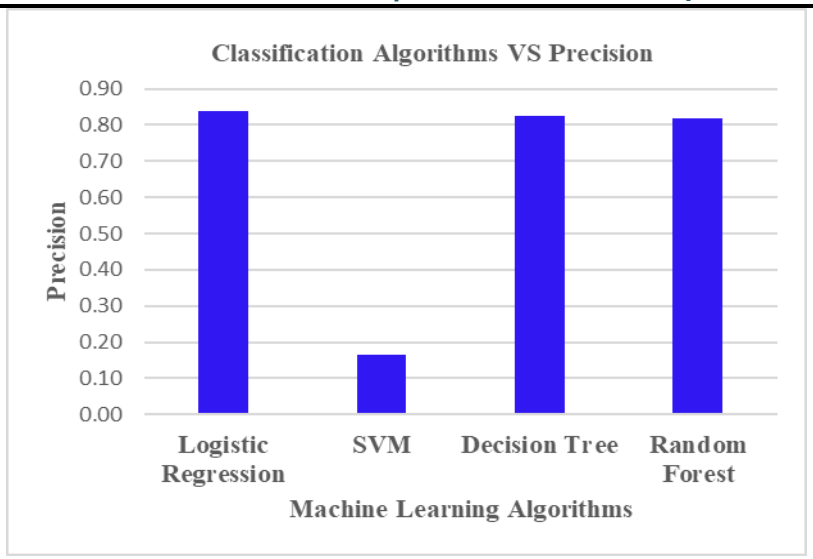

Figure 5. Comparison of Four Algorithms Vs Precision

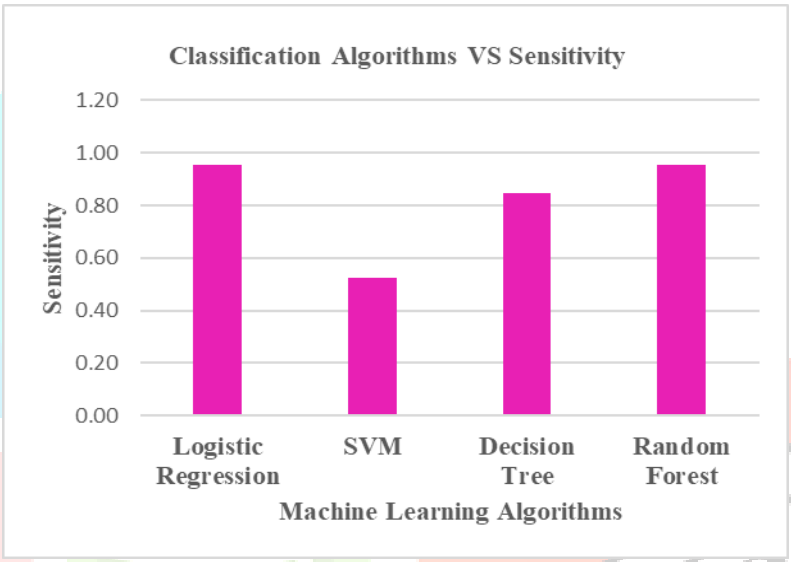

Figure 6. Comparison of Four Algorithms Vs Sensitivity

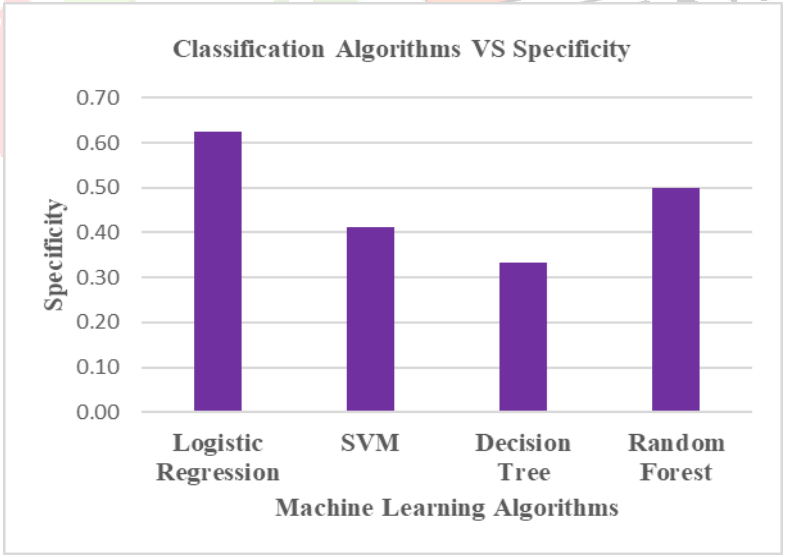

Figure 7. Comparison of Four Algorithms Vs Specificity

VI CONCLUSION

The comparative analysis employed with the machine learning algorithms such as Logistic Regression, Support Vector Machine, Random Forest and Decision Tree. These algorithms are used to predict the liver disease at an early stage. These algorithms were evaluated and compared based on performance metrics such as accuracy, precision, specificity, sensitivity. From the analysis, logistic regression outperforms well than the other algorithms with slight variation and its achieved accuracy is 81.9%. This comparative analysis will help to predict the liver disease and will benefit in managing the health of the individuals.

REFERENCES

[1] https://www.medicinenet.com/liver_disease/article.htm, http://amazingbody.weebly.com/liver.html.
[2] https://www.ncbi.nlm.nih.gov.
[3] Prashant Sahu, Abhishek Bhatt, Anand Chaurasia, Enhanced Hepatoprotective activity of Piperine Loaded Chitosan Microspheres, International Journal of Drug Development & Research, 2012, 4 (4).
[4] http://www.innerbody.com
[5] Standard treatment guidelines and essential medicines list available at www.who.int/selection_medicines
[6] Liver Disease - Pathophysiology of Disease: An Introduction to Clinical Medicine https://accessmedicine.mhmedical.com
[7] Cirrhosis of the liver: Causes, symptoms, and treatments : https://www.medicalnewstoday.com
[8] Bronchitis, emphysema or other - Acute-on-chronic liver failure - NCBI – NIH : https://www.ncbi.nlm.nih.gov
[9] Bendi Venkata Ramana, M. Surendra Prasad Babu, N. B. Venkateswarlu, A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis, International Journal of Database Management Systems, 2011, Vol.3 (2).
[10] Bendi Venkata Raman, M. Surendra Prasad Babu, Liver Classification Using Modified Rotation Forest, International Journal of Engineering Research and Development, 2012,Volume 1(6), 7-24.
[11] Shapla Rani Ghosh and Sajjad Waheed, Analysis of classification algorithms for liver disease diagnosis, J. Sci. Technol. Environ. Inform. 2017, Volume 05 (1), 361-370.
[12] Leoni Sharmila. S, Dharuman. C, Venkatesan. P, Disease Classification Using Machine Learning Algorithms - A Comparative Study, International Journal of Pure and Applied Mathematics 2017, Volume 114 (6), 1-10.
[13] Saranya. A, Seenuvasan. G," A Comparative Study of Diagnosing Liver Disorder Disease Using Classification Algorithm", International Journal of Computer Science and mobile Computing, 2017: 6 (8), 49 - 54.
[14] Banu Priya M, Laura Juliet P. Tamilselvi P.R. Performance Analysis of Liver Disease Prediction Using Machine Learning Algorithms, International Research Journal of Engineering and Technology 2018: 5 (1).
[15] Nazmun Nahar, Ferdous Ara, Liver Disease Prediction by Using Different Decision Tree Techniques, International Journal of Data Mining & Knowledge Management Process, 2018, Vol.8 (2).
[16] Tawseef Ayoub Shaikh, Rashid Ali. Machine Learning: Messiah of 21st Century, Computer Society of India Communications.
[17] Shaik Razia, P. Swathi Prathyusha, N. Vamsi Krishna, N. Sathya Sumana. A review on disease diagnosis using machine learning techniques, International Journal of Pure and Applied Mathematics, 2017, Volume 117(16), 79-85.
[18] Wandra H. K, Mehul Barot, Sarcasm Detection in Sentiment Analysis, International Journal of Current Engineering And Scientific Research, 2017, Volume - 4(9).
[19] https://medium.com/machine-learning.
[20] Indian Liver Patient Dataset available at : http://archive.ics.uci.edu/ml/.
[21] Hossin, M, Sulaiman, M.N, A Review on Evaluation Metrics for Data Classification Evaluations, International Journal of Data Mining & Knowledge Management Process, 2015, Vol.5 (2).