International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

PREDICTION OF LIVER DISEASE USING MACHINE LEARNING

Chauhan Bhavna[1], Assi.Prof.Saket Swarndeep, Dr Gayatri S Pandi[3]

[1]Post Graduate Scholar, Post Graduate Department L J Institute of Engineering and Technology, L J University, Ahmedabad.
[3]Dr Gayatri S Pandi, Head Of Department, Post Graduate Department L J Institute of Engineering and Technology, L J University, Ahmedabad.
[2]Asst. Prof. Post Graduate Department L J Institute of Engineering and Technology, L J University, Ahmedabad.

ABSTRACT

Liver disease is the major cause of death every year. Liver diseases is the fifth big killer in England after cancer, stroke and respiratory disease. The most common causes of liver disease worldwide are chronic hepatitis B and C, alcohol and non alcoholic. Machine Learning has a strong potential in automated diagnosis of various diseases. With the recent upscale in various liver diseases, it is necessary to identify the liver disease at a preliminary stage. In this we propose a new classifier by extending the XGBoost classifier with genetic algorithm. This compares various classification models and visualization techniques used to predict liver disease with feature selection. Outlier detection is used to find out the extreme deviating values and they are eliminated using isolation forest. The performance is measured in terms of accuracy, precision, recall f-measure and time complexity.

Keywords: *Liver disease, Non-alcoholic, fatty liver disease, alcoholic Human life, classification algorithm, Linear Regression, SVM, Machine Learning*

1. INTRODUCTION

The liver is the largest organ in the body, weighing approximately 1.5kg and is vital for life. It is located in the upper right quadrant of the abdomen tucked underneath the ribs. It is responsible for many (over 500) important functions such as processing nutrients, breakdown of toxins and production of essential proteins. It processes digested food from the intestine by breaking it down into glucose and converting into glycogen for energy storage. Another key function is the removal of waste products from the blood. Some are concentrated into bile which is stored in the gallbladder and then is discharged into the duodenum. Bile also plays a part in digestion by emulsifying fats. The liver plays an important role in combating infection and in protein synthesis and metabolism. Chronic liver disease is characterised by **scarring and destruction of the liver tissue**. Early changes, such as 'fatty liver' (a build up of fat in the liver cells) can progress via inflammation (hepatitis) and scarring (fibrosis) to irreversible damage (cirrhosis).

SYMPTOMS

- Skin and eyes that appear yellowish (jaundice)
- Swelling in the legs and ankles.
- Itchy skin.
- Dark urine color.
- Pale stool color.
- Chronic fatigue.
- Nausea or vomiting.

These type of symptoms are seen usually in patients with either TYPE-1 OR TYPE-2 . If any one is having two or more symptoms present in their bodies they have to seek for Medical opinion with specialist doctors without any further delay.

Control of Alcohol Consumption:-

The human liver has to perform around 500 functions every day. A significant amount of alcohol consumption builds fats around the liver known as the alcoholic fatty liver disease, which can belife-threatening. Quitting alcohol consumption or cutting it down as much as one can is the only solution.

Drink Plenty of Water:-

Drinking water is always beneficial. It helps to flush the toxins out of your body and preventsthickening of the blood, thus helps its easy filtering through the liver.

Exercise Regularly:-

One essential part is to start exercising. Non-alcoholic fatty liver disease (NAFLD) is associated with being obese or overweight. Regular exercise will help in preventing liver disease by maintaining a healthy weight. EatA healthy diet is vital to keep the liver healthy. Cut down on thecarbs and eat more nuts. Low carb diet helps in maintaining a healthy liver by controlling the calorie intake.

Cut down on Fizzy Drinks:-

Cold drinks available in the market contain added sugars that lead to obesity. Cut down on theconsumption of such beverages.

2. OBJECTIVES

This paper proposes of a proposed model of machine learning techniques for an efficient prediction of a disease called Liver Disease. Our model will use an improvised and precise technique of detecting the visible patterns of relations between the variables. In the real world patients with Type 2 fatty liver and alcoholic liver disease failing to receive treatment on time which will lead to poor junk food control and increase the risk of complications. If we use machine learning techniques we can reduce the rate of accuracy can be improved. The goal of this work to find an effective manner to find machine learning based model for detecting liver disease in individual clinical data.

3. ALGORITHM STUDIED

Support Vector Machine(SVM):- A Support Vector Machine builds hyperplane or set of hyperplanes in a high dimensional space. It is used for classification as well as regression problems. It chooses the extreme points or vectors that help in creating the hyperplane. As SVM is most effective and precise algorithm among others.

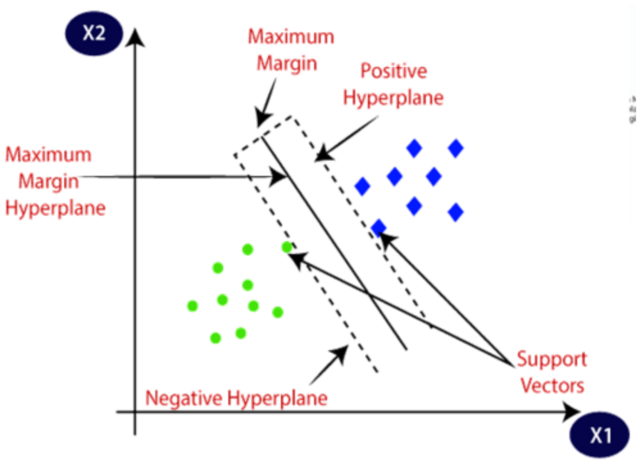


Fig:- 1 Two different categories that are classified using a decision boundary or hyperplane

Navie Bayes:- Navie bayes used for classification and it has high-dimensional training dataset. Navie Bayes classifier is one of the effective algorithm that helps in building machine learning models fast and they can make quick predictions than other algorithms. It is a probabilistic classifier as it predicts on the basis of the probability of the object.

$$\mathrm{P(A\,|\,B)} = \frac{P(B|A)P(A)}{P(B)}$$

Where,

P(A|B) is Posterior probability: Probability of hypothesis A on the observed event B.

P(B|A) is Likelihood probability: Probability of the evidence given that the probability of a hypothesis is true.

P(A) is Prior Probability: Probability of hypothesis before observing the evidence.

P(B) is Marginal Probability: Probability of Evidence.

KNN:- It is based on supervised learning. It is non-parametric algorithm as it doesn't make any assumption on underlying data. KNN algorithm at the process of training dataset it stores the dataset when it gets its new data then it classifies that data into a category that is much similar to the new data.

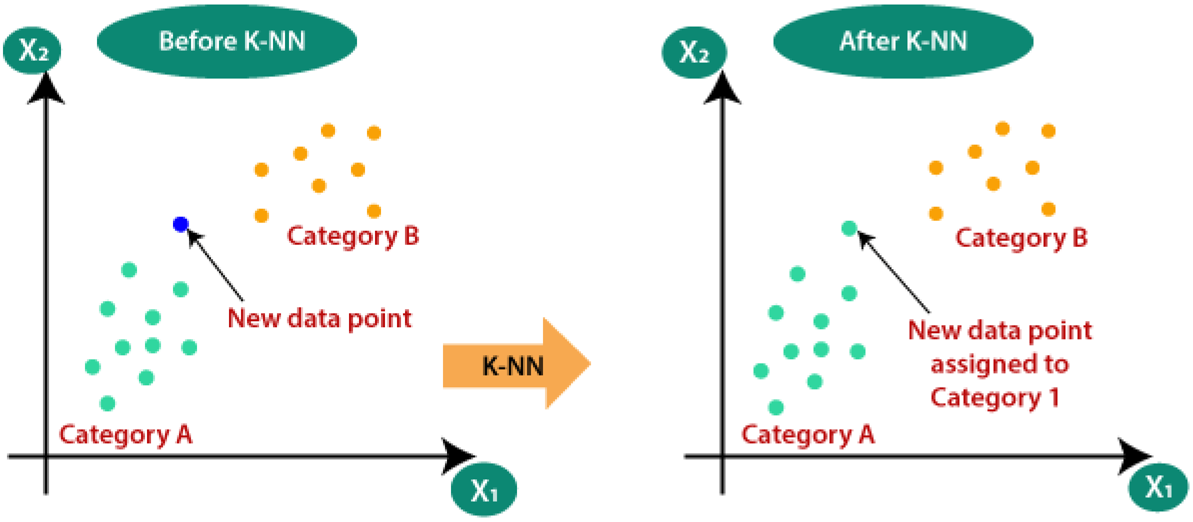


Fig:-2 Classification of KNN according to the new data point.

Artificial Neural Network:- The Term Artificial Neural Network it is a part of biological subfield and it is a part of AI(Artificial Intelligence) that is used to modelled and also develop the structure the brain. Similar to a human brain which has neurons that are interconnected to one another in various layers of the networks. Neurons are weighted that interconnections that regulate the effect of the corresponding to the input signals that is to use of supervised learning to categorize the load parameters of the diabetes.

4. LITERATURE REVIEW

The Motive of this study is to design a model which can detect the liver Disease in patients with maximum accuracy. Dataset is taken for the experiment and in this literature three algorithms are used in this such as naive bayes, decision tree, SVM.

The performance of the algorithm evaluated is measured like accuracy, precision , f-measure and recall. And also result shows that accuracy of naïve bayes is the highest according to this research .

Results are verified using ROC curves in proper and systematic manner.[1]

Previous work on Liver Disease prediction segmentation by which prediction can be given out by different methods mostly they used in dataset to diagnose the liver disease. Data obtained from organization web link are used to analyse the liver disease risk factors applying machine learning techniques. Data transformation improves the performance of the KNN but it doesn't affect the decision tree but it affects the SVM. On the SVM accuracy improved after tuning as before it is approx. 82% and after tuning it increases to 90%. According to this literature three chosen methods but the outcome is the SVM method proves highest accuracy better than other methods.[2]

The aim of this analysis is to develop a system which might predict the hypetyties at high risk level of the patients with a better accuracy. Various information states that machine learning algorithms presents different decision support systems for associating health specialist. The Effectiveness of the decision support system is recognized by its accuracy.[3]

As with the help of machine learning algorithms we have got a flexibility to search an out an answer to the current issue. With the help of machine learning techniques we have been got an advance level of system that has ability to detect and predict the illness with an accurate precision.[3]

This paper discuss the predictive analytics in healthcare. Six different machine learning algorithms are used in this research work. For experimental purpose the dataset is taken that has patient medical report is been listed. ML algorithms are applied and according to their performance and accuracy of the applied algorithm is discussed and compared.[4]

The objective is to study and provide enough understanding to the reader about how healthcare industry can utilize Machine learning algorithm for better decision making and also the prediction of liver disease. Performance analysis of all algorithms are compared in this research work.[4]

This paper has a analytical process is carried out using machine learning algorithms for analysis of the medical data to build the machine learning models to carry out medical diagnosis. This paper explores the approaches to improve the accuracy.

From this literature it is observed that machine learning algorithm takes place in an significant role of knowledge discovery from the database especially in medical diagnosis with the medical data. This paper also predicts higher risk on their disease according to a medical history.[5]

5. RELATED WORK

The aim of this study is to help the patients who are suffering from survey liver disease to help them predicting their disease. We want to help people predicting the disease by developing an system that can bring a change in technology and make it more better and by providing an better treatment and can reduce the cost of the treatment by making it technological way.

The liver disease prediction gives the output from the input variables given to the algorithm this process starts with dataset taken that should be pre processed. Firstly the data is pre processed than the data is cleaned. Removing of the noisy data is done from the huge dataset. Data removal process , data reduction , data dimensionality took place and after that applied algorithm of Machine learning algorithm like naïve bayes, SVM,KNN etc are used . Performance evaluation of all algorithms take place based on various measures like F-Measure, recall, Accuracy, precision. After the performance evaluation comparative analysis is done based on accuracy which algorithm according to the research work has the highest accuracy will be applied to the system. The result predicted output in form of result comes out as an output.

We need to predict the disease of patients disease as there are mainly three types of liver disease and fatty liver as we need to find which is one of them in the patient and also we have to find if any risk factors are there or not in future. As we should know the patients past history as for any side effects and risk factor for future.

After all this process which algorithm is to be used in this system is to be decided. According to the research work mostly SVM has the precise and best accuracy to predict the disease most accurately.

It is evident from the literature survey that the incidence of liver disease is increasing and that although there is evidence that the complication of liver disease can be prevented there are still patients who lack the required knowledge and skills to manage and control their conditions.

This study is an attempt to determine patients and their family members knowledge and views on fatty liver to make recommendations towards improved accuracy education which might lead to improve loyalty to the liver disease treatment .

6. PROPOSED SYSTEM

The main aim is to used the SVM model algorithm. As SVM is a linear model for classification and regression problems. It can solve linear and non-linear problems. The idea of SVM is simple. The algorithm creates a line or hyper plane which separates the data into classes. SVM is an algorithm that takes the data as an input and outputs a line that separates those classes if possible. This is an supervised learning based process. The main parts of process are

Data Collection:- Data Collection is defined as procedure of collecting, measuring and analysing accurate insights for research using standard validant techniques. Data represents information collected in the form of numbers and text. Data collection is generally done after the experiment. Data collection is helpful in planning and estimating. Data collection is qualitative or quantative. During the data collection the researchers must identify the data types, the sources of data types, and what methods are being used.

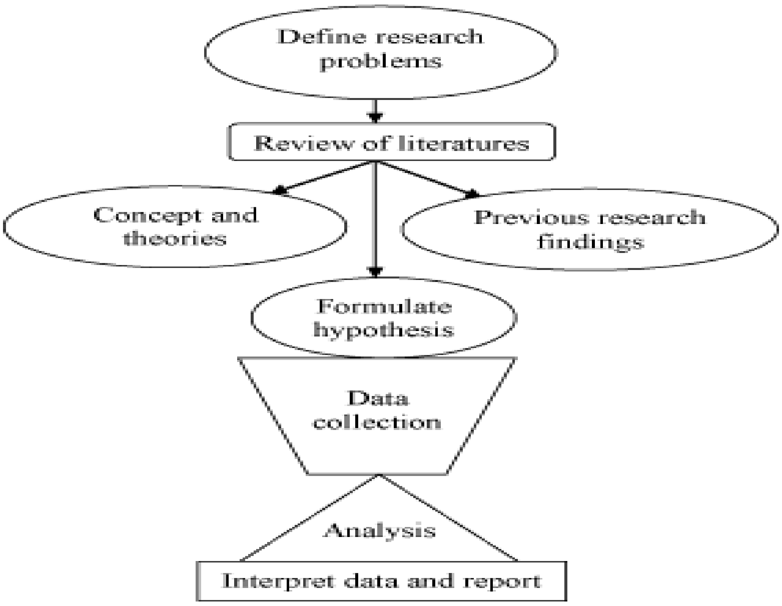


Fig:-4 Review of research process, data collection and analysis.

7. DATA ANALYSIS

Data Analysis is a process used by researchers for reducing data to a story and interpreting it to derive insights. The data analysis process helps in reducing a large chunk of data into a smaller fragments which makes sense. Mostly data analysis are divided into three parts The fist is the data organization. Summarization and categorization together to categorize together contribute to becoming the second known method for data reduction. Third and the last way of data analysis is researchers do it in both top-down or bottom-Up.

Training dataset:- Training dataset is the initial data used to train machine learning models. Training dataset are to machine learning algorithms to teach them how to make predictions or perform the desired task. Training dataset uses machine learning applications to recognize the pattern. As per we have a model of liver disease prediction firstly we will now split the dataset before we train it.

Suppose there are two variables A and B , A will contain all the independent variables and B will contain all dependent variables. After successfully splitting the dataset we will use train_test _split method will be used. As before we build our model we need to see if there are some of zero values in our dataset. We need to check in the head of the dataset if there are zero values in our dataset and zero values are mostly part of independent variables and they have zero values. As because of zero values this can make our model not efficient.

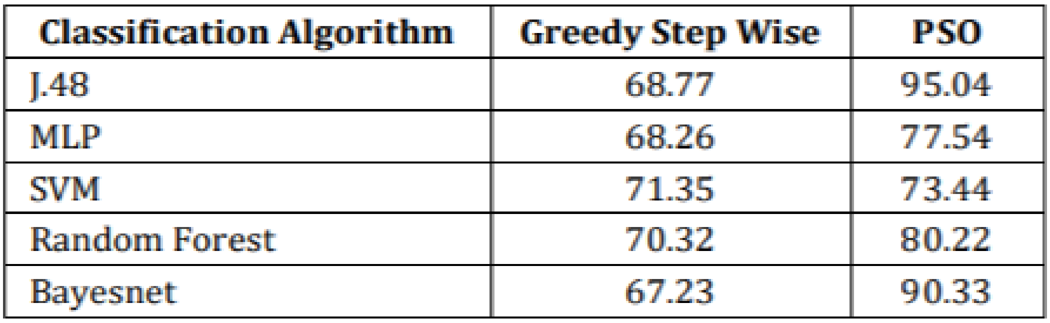
Building The Model:- As stated earlier as we are using the support vector machine method to get the best accuracy score. As accuracy metric is used for the evaluation of model or to evaluate the model. It is the ratio of the correctly predicted probability of instances occurred in a dataset divided by the total number of instances in the dataset. According to the research work support vector machine gives an accuracy score of 0.5321.

Validation Data:- During the training validation data transfers new data into model that has not been evaluated before. Validation data process starts from the first test against unseen data. As it allows the data scientists to evaluate and they may able to know how well the model makes prediction based on new data. As not all data scientist uses the validation data process but it can provide some helpful information to optimize the parameters, which teaches how model should access data. In this sense of validation data occurs as the part of the model training process.

Testing Data:- After the model is built, testing data once again validates that it can make the accurate prediction or not. If training and validation data include labels to monitor performance metrices of the model. The testing data should be unlabelled. Test data provides an final, real-world check of an unseen dataset to confirm that the machine learning algorithm war trained effectively.

Accuracy:

The following table describes an overall classification algorithm for accuracy values analysis. In this table contains existing and proposed accuracy a values shows, Accuracy = (TP + TN) / (TP + TN + FP+ FN)



Classification Algorithm	Greedy Step Wise	PSO
J.48	68.77	95.04
MLP	68.26	77.54
SVM	71.35	73.44
Random Forest	70.32	80.22
Bayesnet	67.23	90.33

Figure 2.2 Accuracy table

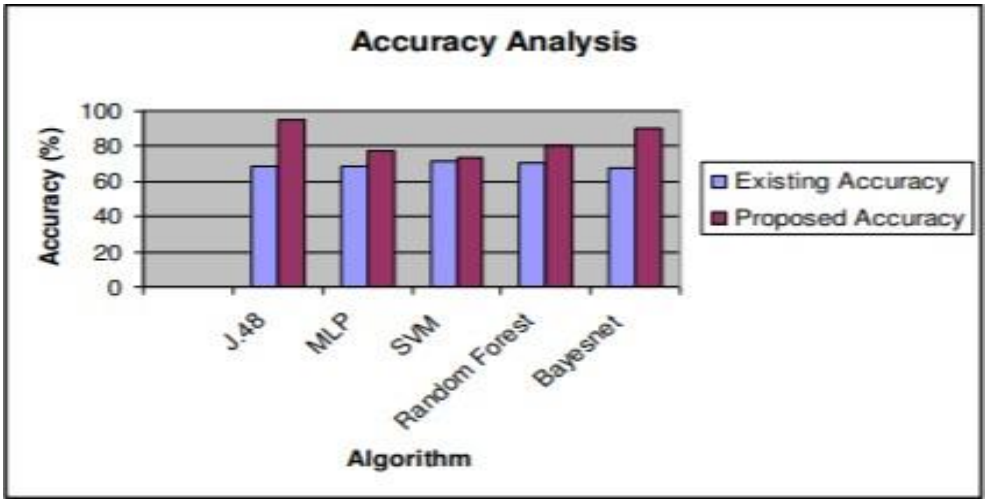


Figure 2.3 Accuracy and Consumption Time

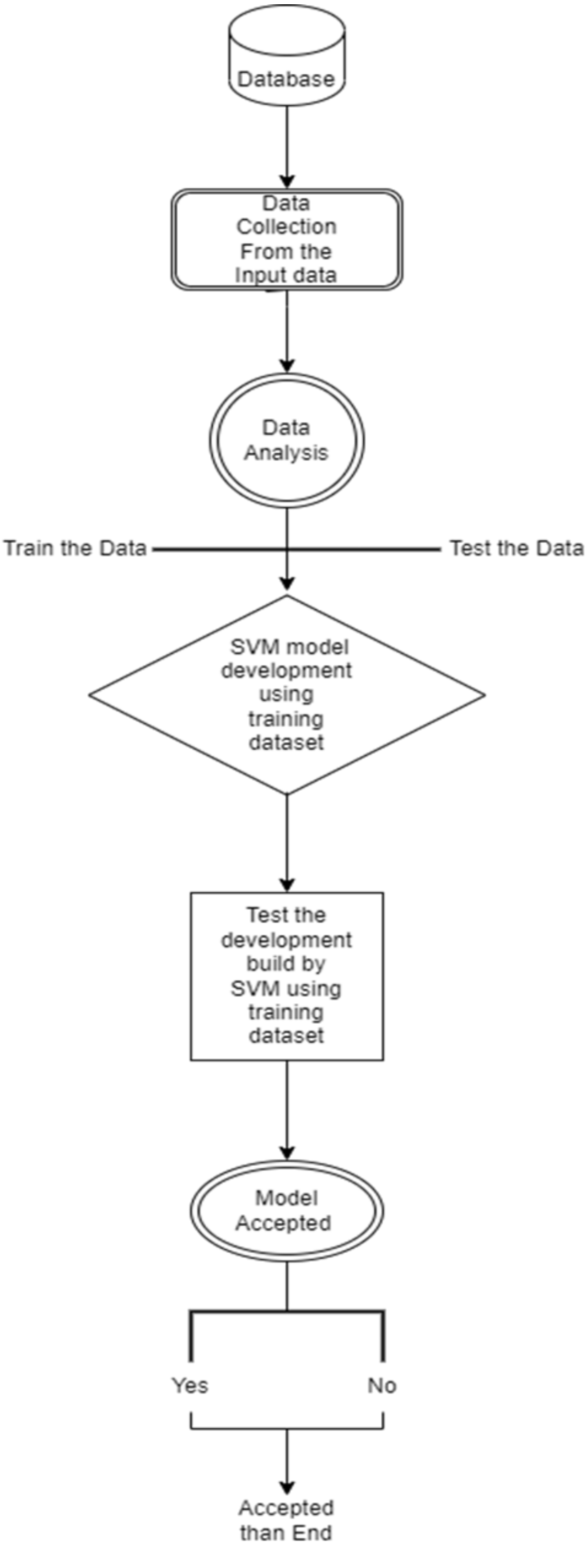
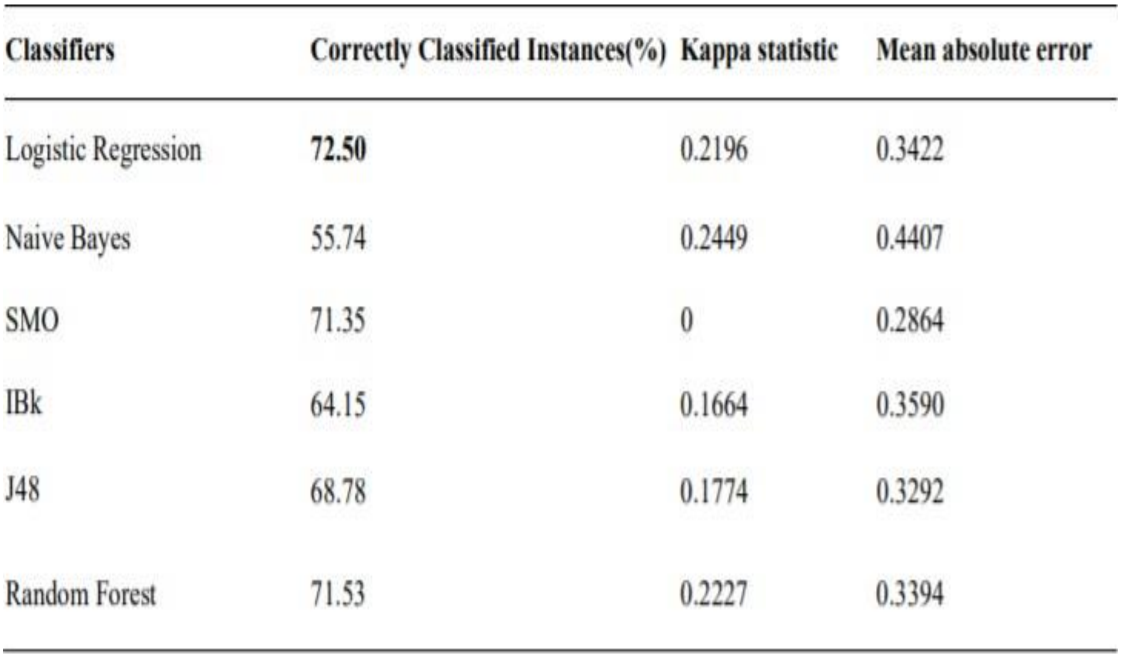
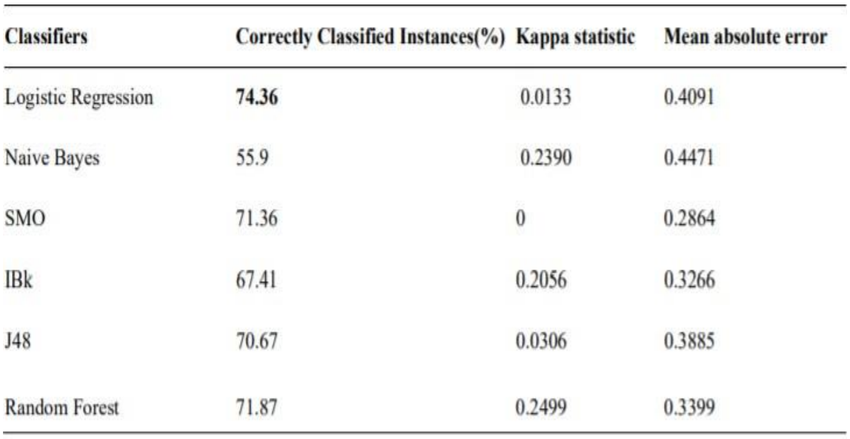

Database
Data Collection From the Input data
Data Analysis
Train the Data
Test the Data
SVM model development using training dataset
Test the development build by SVM using training dataset
Model Accepted
Yes
No
Accepted than End

Fig:- 5 Proposed System Of Support Vector Machine

Compare the results of different classifiers using feature selection technique onLiver Patient Diseases Dataset



Classifiers	Correctly Classified Instances(%)	Kappa statistic	Mean absolute error
Logistic Regression	**72.50**	0.2196	0.3422
Naive Bayes	55.74	0.2449	0.4407
SMO	71.35	0	0.2864
IBk	64.15	0.1664	0.3590
J48	68.78	0.1774	0.3292
Random Forest	71.53	0.2227	0.3394



Classifiers	Correctly Classified Instances(%)	Kappa statistic	Mean absolute error
Logistic Regression	**74.36**	0.0133	0.4091
Naive Bayes	55.9	0.2390	0.4471
SMO	71.36	0	0.2864
IBk	67.41	0.2056	0.3266
J48	70.67	0.0306	0.3885
Random Forest	71.87	0.2499	0.3399

Table patient dataset result

The above graph shows the comparison of different classifiers accuracy percentage using featureselection and without using feature selection techniques.

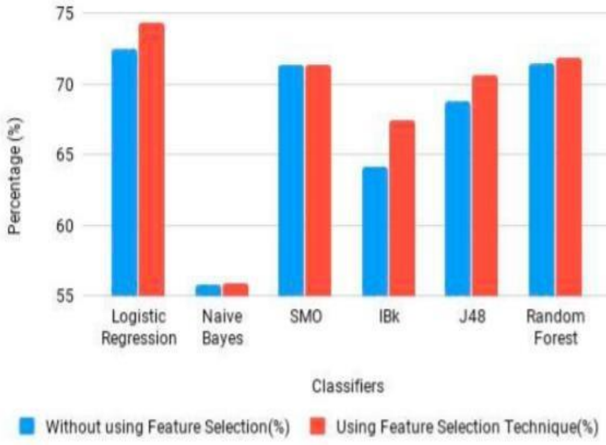


Figure: - Finalize Number of Iteration Result [7]

Feature Importance:-

Feature Importance refers to techniques that calculate a score for all the input features for a given model. The score only represents the importance of the feature. As much is the score is higher of that particular specific feature will have higher effect on the model that is being used to predict a certain Variable. Feature importance works in way that it will rank features based on the effect that they have on their model prediction.

There are different ways to calculate feature importance according to this research paper we have studied mainly two methods Gini importance and permutation feature importance. According to the scikit-learn Gini importance is used to calculate the node impurity and feature importance is basically reduction of number of samples in the impurity of node that are weighted by the number of samples that are reaching that node from the total number of samples. The equation of two nodes is:-

$$ni_j = w_j C_j - w_{left(j)} C_{left(j)} - w_{right(j)} C_{right(j)}$$

A single feature can be used in the different branches of the tree. The features are normalized against all the sum of features values present in the tree.

$$fi_i = \frac{\sum_{j:\text{node j splits on feature i}} ni_j}{\sum_{j \in \text{all nodes}} ni_j}$$

Now as we talk about permutation feature importance the idea behind permutation feature importance is simple. The feature importance is calculated by increase or decrease in the error when we permute the values of feature. The best thing about this method is that it can be applied to every machine learning model.in this feature importance methods there are no complex mathematical formulas behind it. The permutation feature importance is based on an algorithm that works on this factor as follows:-

1. Calculate the mean squared error with the original values.
2. Shuffle the values for the features and make predictions.
3. Calculate the mean squared error with the shuffled values.
4. Compare the differences between them.
5. Sort the differences in descending order to get features with most to least importance.

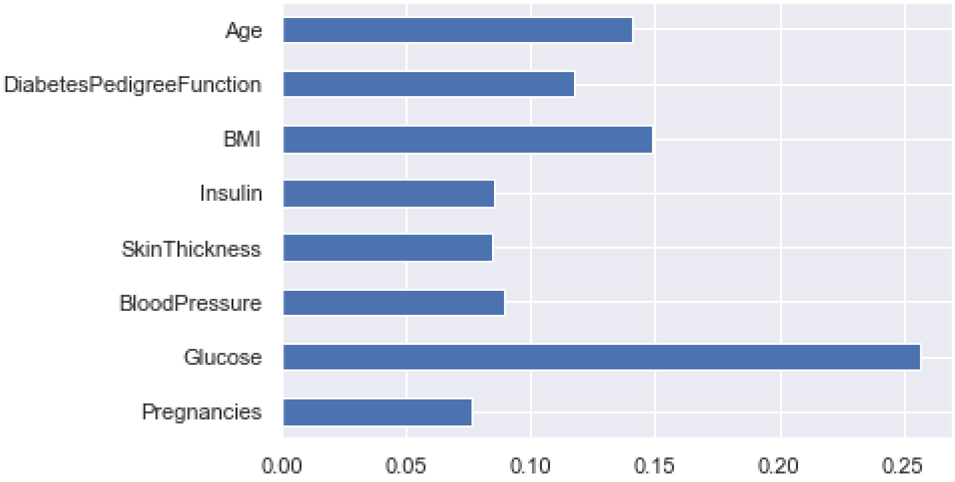


Fig:--6 Feature Importance

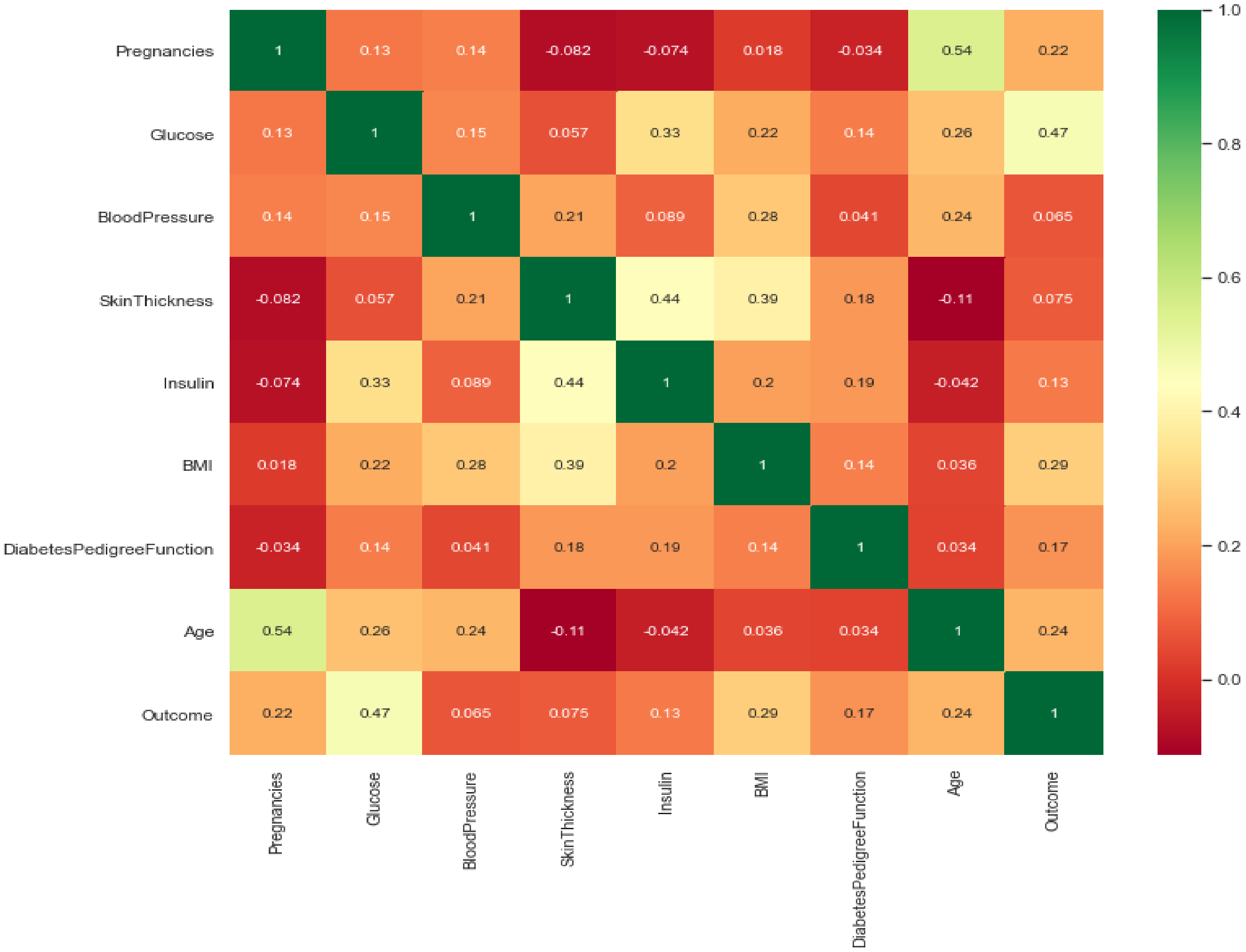


Fig:- 7 confusion Matrix

8. PROBLEM STATEMENT

Patients with Liver disease have been continuously increasing because of excessive consumption of alcohol, inhale of harmful gases, intake of contaminated food, pickles and drugs. This dataset was used to evaluate prediction algorithms in an effort to reduce burden on doctors.It is a binary classification problem, where given the above set of features, we need to predict if a given patient has liver disease or not

9. CONCLUSION

This Study will help you to know which gives the best accuracy among all algorithm that gives the precise result. This is the one of the most important medical problem is the liver disease and it is very rare and it may be diagnose at an early stage. In this study or research systematic efforts are made for the user to detect the disease at an early stage and we can take precautions and take pecuniary measures and solve the problem. Algorithm like SVM, Naive Bayes classifier, Decision tree using machine learning algorithm are used in this system. Experimental outcomes are there according to their algorithms and with the help of the accuracy, precision etc.

REFERENCES

[1] Ailem, M., Role, F. and Nadif, M., 2015, October. Co-clustering document-term matrices bydirectmaximization of graph modularity. In Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (pp. 1807-1810). ACM.

[2] Ailem, M., Role, F. and Nadif, M., 2017. Sparse poisson latent block model for document clustering. IEEE Transactions on Knowledge and Data Engineering, 29(7), pp.1563-1576.

[3] Liao, K., Liu, G., Xiao, L. and Liu, C., 2013. A sample-based hierarchical adaptive K-means clustering method for large-scale video retrieval. Knowledge-Based Systems, 49, pp.123- 133.

[4] Onan, A., Bulut, H. and Korukoglu, S., 2017. An improved ant algorithm with LDA-based representation for text document clustering. Journal of Information Science, 43(2), pp.275- 292.

[5] Rossi, R.G., Marcacini, R.M. and Rezende, S.O., 2013. Benchmarking text collections for classification and clustering tasks. Institute of Mathematics and Computer Sciences, University ofSao Paulo.

[6] Jain, A.K., 2010. Data clustering: 50 years beyond K-means. Pattern recognition letters, 31(8), pp.651-666.

[7] Park, H.S. and Jun, C.H., 2009. A simple and fast algorithm for K-medoids clustering. Expert systems with applications, 36(2), pp.3336-3341.

[8] Neha, D. and Vidyavathi, B.M., 2015. A survey on applications of data mining using clustering techniques. International Journal of Computer Applications, 126(2).

[9] Djenouri, Y., Belhadi, A. and Belkebir, R., 2018. Bees swarm optimization guided by data mining techniques for document information retrieval. Expert Systems with Applications, 94,pp.126-136.

[10] Reddy, G.S., Rajinikanth, T.V. and Rao, A.A., 2014, February. A frequent term based text clustering approach using novel similarity measure. In 2014 IEEE International Advance Computing Conference (IACC) (pp. 495-499). IEEE.

[11] Alhawarat, M. and Hegazi, M., 2018. Revisiting K-Means and Topic Modeling, a Comparison Study to Cluster Arabic Documents. IEEE Access, 6, pp.42740-42749.

[12] Chen, Y. and Sun, P., 2018, August. An Optimized K-Means Algorithm Based on FSTVM. In 2018 International Conference on Virtual Reality and Intelligent Systems (ICVRIS) (pp. 363- 366). IEEE.

[13] Janani, R. and Vijayarani, S., 2019. Text document clustering using Spectral Clustering algorithm with Particle Swarm Optimization. Expert Systems with Applications, 134, pp.192-200.

[14] Jin, C.X. and Bai, Q.C., 2016, June. Text clustering algorithm based on the graph structures of semantic word co-occurrence. In 2016 International Conference on Information System and Artificial Intelligence (ISAI) (pp. 497-502). IEEE.

[15] Afzali, M. and Kumar, S., 2019, February. Text Document Clustering: Issues and Challenges. In 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon) (pp. 263-268). IEEE.

[16] H. Bunke, M. Roth and E.G.Schukat-Talamazzini, “Offline Cursive Handwriting Recognition Using Hidden Markov Models”, Pattern Recognition, Vol. 28, No. 9, 1995 Elsevier Science Ltd.

[17] Nafiz Arica, “An Off-line Character Recognition System for free style Handwriting”, a thesis submitted to the graduate school of natural and applied sciences of the middle east technical university,1998.

[18] Yong Haw Tay, Pierre-Michel Lallican, Marzuki Khalid, Christian Viard-Gaudin, Stefan Knerr,” An Offline Cursive Handwritten Word Recognition System”, IEEE Catalogue No. 01 CH37239 2001.

[19] Anshul Gupta, Manisha Srivastava and Chitralekha Mahanta,” Offline Handwritten Character Recognition International Conference on Computer Applications and Industrial Electronics (ICCAIE), 2011.

[20] J. Pradeep, E. Srinivasan and S. Himavathi,”Neural Network Based Recognition System Integrating Feature Extraction and Classification for English Handwritten”, International Journal of Engineering (IJE) Transactions B: Applications Vol. 25, No. 2, (May 2012) 99-106.

[21] D. K. Patel, T. Som and M. K Singh,” Improving the Recognition of Handwritten Characters using Neural Network through Multiresolution Technique And Euclidean Distance Metric”, International Journal of Computer Applications (0975 – 8887) Volume 45– No.6 May2012.

[22] M. Blumenstein, B. Verma and H. Basli,” A Novel Feature Extraction Technique for the Recognition of Segmented Handwritten Characters”, Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR’03) 0-7695-1960-1/03 $17.00 © 2003 IEEE

[23] Sumedha B. Hallale, Geeta D. Salunke, ” Twelve Directional Feature Extraction for Handwritten English Character Recognition”, International Journal of Recent Technology and Engineering (IJRTE)ISSN:2277-3878, Volume-2, Issue-2, May 2013.

[24] Amit Choudhary, Rahul Rishi and Savita Ahlawat, ”Off-Line Handwritten Character Recognition using Features Extracted from Binarization Technique”, 2212-6716 © 2013 American Applied Science Research Institute doi:10.1016/j.aasri.2013.10.045

[25] Rafael M. O. Cruz, George D. C. Cavalcanti and Tsang Ing Ren,” An Ensemble classifier for offline Cursive character recognition using multiple features Extraction technique”, 978-1- 4244-8126- 2/10/$26.00 ©2010 IEEE

[26] Singh, Sameer, Mark Hewitt,“Cursive Digit And Character Recognition on CedarPattern Recognition, 2000.Proceedings. 15th international conference on. Vo.

[27] K. Gaurav and Bhatia P. K., “Analytical Review of Preprocessing Techniques for Offline Liver disease predictoin”, 2nd International Conference on Emerging Trends in Engineering &Management, ICETEM, 2013.

[28] Charleonnan,T. Fufaung, T. Niyomwong, W Chokchueypattanakit, S. Suwannawach, N. Ninchawee “Predictive Analytics for Chronic Kidney Disease Using Machine Learning Techniques” *MITiCON-2016.*Anatomy and function of the liver.

[29] Abhishek Chowdhury, Thirunavukkarasu K, Sidhyant Tejas(2017), Predicting whether song will behit using Logistic Regression. Volume 6 Issue 9 September 2017.

[30] K-Nearest Neighbours. [Online]Available: https://www.saedsayad.com/k_nearest_neighbors.htm P.Mazaheri, A. Narouziand A. Karimi (2015), Using Algorithms to Predict Liver, Electronics.