Identification and Prediction of Liver Disease using Logistic Regression

[1]Neeraj Varshney, [2]Ashish Sharma,

[1]Neeraj Varshney, [2]Ashish Sharma,
Department of Computer Engineering, GLA University, Mathura.
Department of Computer Engineering and Application, GLA University, Mathura
E-Mail: neeraj.varshney@gla.ac.in, ashishs.sharma@gla.ac.in

History:

- Received:
- Accepted:
- Published:

ABSTRACT

Identification of disease at a beginning stage is very essential for higher treatment. It's a awfully complicated task for medical researchers to predict the illness within the early stages because of delicate symptoms. Typically the symptoms turn out to be evident once it's too late. to beat this issue, this project aims to boost disease designation victimization machine learning approaches. The most objective of this analysis is to use categorization techniques to spot the liver patients from healthy people. This project conjointly aims to match the categorization techniques supported their presentation factors. To serve the medical community for the designation of disease between patients, a graphical computer interface is urbanized victimization python (Node RED). The GUI will be promptly used by doctors and medical practitioners as a screening tool for the disease.

Keywords: Liver Disease, Machine Learning, Classification, Node RED, GUI

Introduction

The Liver, which is the biggest strong organ in the human body, plays out a few significant capacities. Its significant capacities incorporate assembling fundamental proteins and blood thickening components, processing fat and starches, and so on.

Exorbitant utilization of liquor, infections, the admission of sullied nourishment and tranquilizes, etc are the significant reasons for liver sicknesses. The indications could possibly be unmistakable in the beginning periods. If not analyzed in the underlying stages, liver maladies can prompt hazardous conditions.

In India, as indicated by the most recent information distributed by WHO in May 2014, liver infection passing comprises to 2.44% of all out passing's. Additionally, around 10 lakh individuals are determined to have liver maladies consistently in India. With expanded level of populace falling prey to this ailment, it is basic to see approaches to identify these disarranges in beginning periods. This not just spares a human life, yet in addition with the approach of innovation, it makes the clinical treatment increasingly exact and quicker. Furthermore, the expansion in the quantity of cases adds to the structure of database while applying innovation.

Classification is a viable method used to deal with such issues in the clinical field. Utilizing the accessible component esteems, the classifier could foresee whether an individual has liver illness. This capacity will assist specialists with recognizing the sickness ahead of time. It is always recommended to reduce Type I error due to the rejection of null hypothesis (as false) when it is actually true), as false diagnosis could lead to fatal conditions.

Related Works

Most importantly writing's and research papers were looked into for getting more data about the issue and realizing which kind of work was finished by others on this theme and by which strategy.

Hoon Jin et al. [2], depicted the idea of a variety of categorization techniques that help the specialists to decide the disease rapidly and effectively. Different classifiers, for example, k-NN, Naïve Bayes, Multi-Layer Perceptron and Decision Tree were thought about and broke down dependent on a few parameters like sensitivity, specificity and so on. Algorithms had implemented utilizing the Weka instrument and from UCI Repository dataset was gathered. The exploratory outcomes indicated that as far as accuracy, Naïve Bayes returns better order results though Random Forest and Logistic Regression gave better outcomes regarding sensitivity and recall.

Dr. S. Vijayarani1 et al. [3], shown the predictive analysis of liver issue utilizing different classification algorithms. Right now, Bayes and Support Vector Machine grouping calculations were utilized. These two calculations were looked at based on execution parameters that incorporate order execution time measures and exactness measures. The proposed framework was actualized utilizing Matlab 2013 apparatus and from UCI Repository the dataset had been gathered. After the trial results, it had been seen that Naïve Bayes Algorithm beaten by Support Vector Machine because of the most elevated categorization exactness and can be utilized further in the forecast of liver illness.

METHODOLOGY

This project aims to predict whether the patient has liver disease or not. The model is provided with a dataset [1] that contains the blood test values of liver patients in India. The Architecture delineated in the accompanying figure, has been embraced for leading liver patient dataset classification experiment.

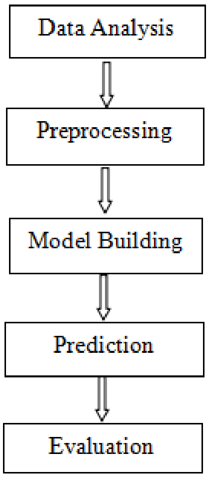


Fig. 1 **Architecture**

A. Data Analysis and Pre-processing

This is general looking at the data to figure out what is going on. Data Analysis is a short hand for getting data, reshaping it, exploring it and visualizing it. Data Preprocessing is a technique which converts the raw data into acceptable format. It will check whether there is any missing data, irrelevant data and fill the absent values by hand or by mean or by considering most probable value.

B. Model Building

The objective right now the procedure is to pick a calculation and it is connected with the sort of issue we are confronting (Regression or Classification).For our model we are confronting Classification Problem. For this we utilized KNN, Logistic Regression, Decision Tree, SVM, Random Forest algorithms.

1) Logistic Regression: Logistic regression is primarily a random forest algorithm. The output y, can take distinct values of inputs, X in classification problem.

In spite of main stream thinking, it is a regression model. A regression model is built by the model to predict the probability that a given input belongs to numerical category "1". Sigmoid function is used to model data by logistic regression whereas linear regression surmises that the input go along with a linear function.

$$g(z) = \frac{1}{1+e^{-z}}$$

Fig. 2 **Sigmoid Function**

2) Support Vector Machines: In ML, SVMs conjointly called support vector networks area unit supervised learning models with connected learning algorithms that examine input used for regression analysis and classification.

A Support Vector Machine could be a selective classifier properly characterized by a filtering hyper plane. In different quarters, for given known coaching input (supervised learning), a wonderful hyper plane that defines inchoate examples is that the output for the algorithmic program.

An SVM model could be a depiction of the instances as spots in area, aforethought so the instances of the distinct sorts area unit separated by a transparent area that's as broad as viable.

Along with death penalty linear classification, SVMs will like an expert accomplish a non-linear classification, fully depict their information into high-dimensional detail gaps.

3) K-Nearest Neighbor: When we consider classification algorithms in machine learning, KNN is the utmost rudimentary yet essential one. It finds profound application in intrusion detection, data mining and pattern recognition and it is in the hands of supervised learning province.

As it is distribution free which means it does not make any fundamental expectation about the dispensation of input (in opposition to other algorithms such as GMM, which suppose a Gaussian issue of the given input) it is broadly replaceable in real-life scenarios.

4) Decision Tree: The most potent and common implement for regression and classification in machine learning is call tree. A call tree could be a tree structure like flow diagram, wherever take a look at on a credit is denoted by internal node and tests outcome is delineated by every branch and sophistication label is hold by every leaf node or terminal node.

By sorting instances down the tree from the root node to leaf node, to produce the classification of the instance call trees reason instances. Associate in Nursing example is classified by starting at the foundation node of the tree, testing the attribute determined by this node, then moving down the tree limb admire the estimation of the attribute. This method is then duplicated for the sub tree embedded at the new node.

5) Random Forest: Random forest is an AI relapse technique for characterization that drive by building liver info into an out sized range of alternative trees at making ready time and yielding the category that's the strategy of the categories yield by singular trees . it's unsurpassed in accuracy among current calculations. It yields arrangement effectively on vast liver knowledge set. It will influence an out sized

range of data attributes while not variable cancellation. It offers assessments of what factors are important within the order. Impulsive Forests develops various arrangement trees. To cluster on other liver article from an information vector, place the knowledge vector down all of the trees within the ground. Every tree offers associate degree order, and says the tree "votes" for that category. The ground picks the grouping having the foremost votes).

C. Training and Prediction

The training procedure includes introducing some arbitrary qualities for every one of the preparation frameworks and endeavors to anticipate the yield of the info information utilizing the underlying irregular qualities. Toward the start, the blunder will be enormous, however contrasting the model's forecast and the right yield, the model can change the loads and inclinations esteems until having a decent anticipating model. By and large 80% of the information is utilized for training and 20% is utilized for Testing.

D. Evaluation

Classification Accuracy, Area under Curve and Confusion Matrix are used as evaluation metrics.

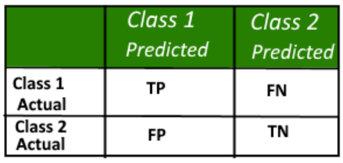
1) Classification Accuracy: Classification Accuracy is that the issue that we have a tendency to frequently mean, after we utilize the term exactness. It's the proportion of variety of right predictions to the whole variety of input samples.

The formula for classification accuracy is as follows:

$$Accuracy = \frac{Number\ of\ Correct\ predictio}{Total\ number\ of\ predictions\ m}$$

Fig. 3 Accuracy

2) uncertainty Matrix: uncertainty Matrix because the name suggests provides us a matrix as output and describes the whole performance of the model.



	Class 1 Predicted	Class 2 Predicted
Class 1 Actual	TP	FN
Class 2 Actual	FP	TN

Fig. 4 uncertainty Matrix

Here,

Class1: Positive

Class2: Negative

TP: Both prediction and observation are positive.

FN: Prediction is negative but observed is positive.

TN: Both observation and predicted are negative.

FP: Prediction is positive but observed is negative.

3) Area under Curve: Area under Curve (AUC) is one among the foremost generally utilized measurements for assessment. It's utilized for binary grouping issue. AUC of a classifier is likelihood the probability that the classifier can rank a arbitrarily picked positive model more than an arbitrarily picked negative example. Before process AUC, let us to comprehend 2 elementary terms:

a. True Positive Rate: It is additionally referred to as Sensitivity. It's outlined as TP/ (FN+TP). True Positive Rate corresponds to the proportion of positive information points that area unit properly thought-about as positive, with reference to all positive information points.

$$TruePositiveRate = \frac{TruePositive}{FalseNegative + TruePositive}$$

Fig. 5 True Positive Rate

b. False Positive Rate:It is additionally referred to as Specificity. It's outlined as FP / (FP+TN). False Positive Rate corresponds to the proportion of negative information points that area unit erroneously thought-about as positive, with reference to all negative information points.

$$FalsePositiveRate = \frac{FalsePositive}{FalsePositive + TrueNegative}$$

Fig. 6 False Positive rate

RESULTS & DISCUSSION

Accuracy values of classification algorithms we used like KNN, Logistic Regression, SVM, Decision Tree Classifier and Random Forest are displayed in below table. Out of all these algorithms Logistic regression algorithm got the highest Accuracy.

TABLE I ACCURACY

Algorithm	Accuracy
Logistic Regression	71.42
K-NN	64
SVM	70.28
Decision Tree	64.57
Random Forest	66.85

The below figure shows the accuracy values of classification algorithms we used.

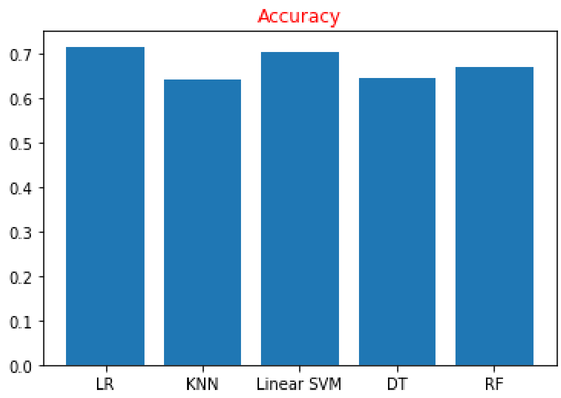
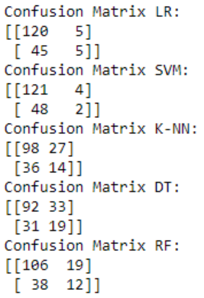


Fig. 7 **Accuracy of Algorithms**

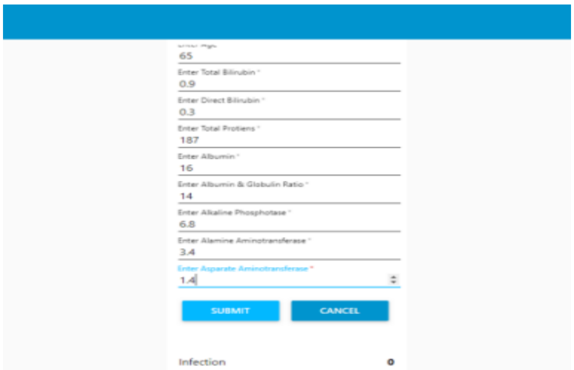
The below figure shows the confusion matrix of the algorithms we used.



```
Confusion Matrix LR:
[[120   5]
 [ 45   5]]
Confusion Matrix SVM:
[[121   4]
 [ 48   2]]
Confusion Matrix K-NN:
[[98 27]
 [36 14]]
Confusion Matrix DT:
[[92 33]
 [31 19]]
Confusion Matrix RF:
[[106  19]
 [ 38  12]]
```

Fig. 8 **Confusion Matrix of Algorithms**

The below figure describes that the user interface after reading all the fields values, predicting that the patient has no liver disease.

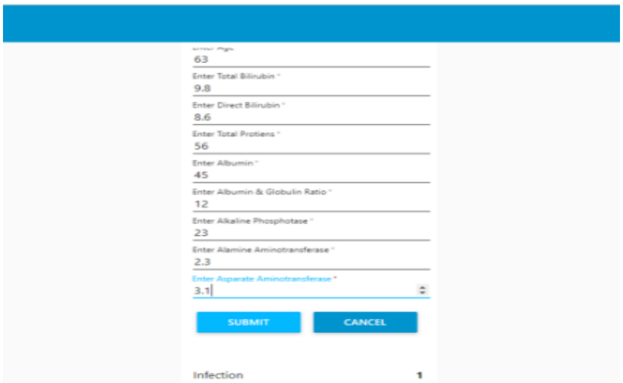


65
Enter Total Bilirubin *
0.9
Enter Direct Bilirubin *
0.3
Enter Total Protiens *
187
Enter Albumin *
16
Enter Albumin & Globulin Ratio *
14
Enter Alkaline Phosphotase *
6.8
Enter Alamine Aminotransferase *
3.4
Enter Asparate Aminotransferase *
1.4

Infection 0

Fig. 9 **UI predicted as not liver patient**

The below figure describes that the user interface after reading all the fields values, predicting that the patient has liver disease.



63
Enter Total Bilirubin *
9.8
Enter Direct Bilirubin *
8.6
Enter Total Protiens *
56
Enter Albumin *
45
Enter Albumin & Globulin Ratio *
12
Enter Alkaline Phosphotase *
23
Enter Alamine Aminotransferase *
2.3
Enter Asparate Aminotransferase *
3.1

Infection 1

Fig. 10 **UI predicted as liver patient**

Conclusion

In this project, diagnosing of liver disease in patients is done by using machine learning algorithms. Logistic Regression, Support Vector Machines, Decision Tree Classifier, KNN and Random Forest are the machine learning techniques we were used. The system was implemented by evaluating the presentation of all the representations used. Based on the patient records we determine whether the patient has liver disease or not. This representation utilized to reduce burden on doctors.

FUTURE WORK

Right now blood test reports specialists need to enter the fields of UI manually. To beat this issue there is need to build up an IOT application that naturally enters the blood test esteems in UI and get whether the patient has liver infection or not.

REFERENCES

1. P. Rajeswari ,G. Sophia Reena , Analysis of Liver Disorder Using Data Mining Algorithm,Global Journal of Computer Science and Technology,2010.
2. Sa'sdiyah Noor Novita Alfisahrin,Teddy Mantoro, Data mining Techniques For Optimatization of Liver Disease Classification,International conference on advanced Computer Science Application and Technologies,2013.
3. S. Dhamodharan , Liver Disease Prediction Using Bayesian Classification , National Confrence on Advanced Computing,Application&Technologies,2014
4. S.E.Sekar ,Y.Unal, Z.Erdem,and H.Erdinc Kocer,Ensembled Correlation Between Liver Analysis Output, International Journal of Biology and Biomedical ngineering,ISSN:1998-4150
5. A.S.Aneesh kumar,Dr.C.Jothi Venkateswaran , A novel approach for Liver disorder Classification using Data Mining Techniques ,Engineering and Scientific International Journal ,ISSN 2394- 7179,ISSN 2394-7187,2015. Hoon Jin, Seoungcheon Kim, Jinhong Kim, —Decision Factors on Effective Liver Patient Data Predictionǁ, International Journal of BioScience and Bio-Technology, Vol. 6, Issue.4, pp. 167-178, 2014.
6. P. Thangarajul,R.Mehala, Performance Analysis of PSO-KStar Classifier over Liver Diseases,

International Journal of Advanced Research in Computer Engineering, 2015.

7. Onwodi Gregory, Prediction of Liver Disease (Biliary Cirrhosis) Using Data Mining Technique, International Journal of EmergingTechnology&Research, ISSN (E):2347-5900, ISSN (P):2347-6079, 2015.
8. Dr.S.Vijayarani,Mr.S.Dhayanand,Liver Disease Prediction using SVM and Naïve Bayes Algorithms, International Journal of Science, Engineering and Technology Research(IJSETR), 2015.
9. Ebenezer Obaloluwa Olaniyi khashman Aadnan, "Liver DiseaseDiagnosisBasedon Neural Networks" , Advances in Computational Intelligence, Proceedings of the 16th International Conference on Neural Networks (NN '15), November 7-9, 2015.
10. Anju Gulia, Dr. Rajan Vohra, Praveen Rani, "Liver Patient Classification UsingIntelligent Techniques", International Journal of Computer Science and Information Technologies, Vol. 5 (4), 2014, 5110-5115. [13] Dash M., Liu H., "Feature Selection for Classification," Intelligent Data Analysis, Elsevier, pp. 131 -156, 1997
11. Dr. S. Vijayarani, Mr.S.Dhayanand, ―Liver Disease Prediction using SVM and Naïve Bayes Algorithms‖, International Journal of Science, Engineering and Technology Research (IJSETR), Vol. 4, Issue.4 pp. 816-820, 2015.
12. Kumar, Manoj, and Ashish Sharma. "Mining of data stream using "DDenStream" clustering algorithm." 2013 IEEE International Conference in MOOC, Innovation and Technology in Education (MITE). IEEE, 2013.
13. Sharma, Ashish, Anant Ram, and Archit Bansal. "Feature Extraction Mining for Student Performance Analysis." Proceedings of ICETIT 2019. Springer, Cham, 2020. 785-797.
14. Sharma, Ashish, and Dhara Upadhyay. "VDBSCAN Clustering with Map-Reduce Technique." Recent Findings in Intelligent Computing Techniques. Springer, Singapore, 2018. 305-314.
15. Sharma, Ashish, Ashish Sharma, and Anand Singh Jalal. "Distance-based facility location problem for fuzzy demand with simultaneous opening of two facilities." International Journal of Computing Science and Mathematics 9.6 (2018): 590-601.
16. Agarwal, Rohit, A. S. Jalal, and K. V. Arya. "A review on presentation attack detection system for fake fingerprint." Modern Physics Letters B 34.05 (2020): 2030001.
17. Mishra, Ayushi, et al. "A robust approach for palmprint biometric recognition." International Journal of Biometrics 11.4 (2019): 389-408.
18. Singh, Anshy, Shashi Shekhar, and Anand Singh Jalal. "Semantic based image retrieval using multi-agent model by searching and filtering replicated web images." 2012 World Congress on Information and Communication Technologies. IEEE, 2012.
19. Shekhar, Shashi, et al. "A WEBIR crawling framework for retrieving highly relevant web documents: evaluation based on rank aggregation and result merging algorithms." 2011 International Conference on Computational Intelligence and Communication Networks. IEEE, 2011.
20. Varun K L Srivastava, N. Chandra Sekhar Reddy, Dr. Anubha Shrivastava, "An Effective Code Metrics for Evaluation of Protected Parameters in Database Applications", International Journal of Advanced Trends in Computer Science and Engineering, Volume 8, No.1.3, 2019. doi.org/10.30534/ijatcse/2019/1681.32019
21. Varun K L Srivastava , N. Chandra Sekhar Reddy , Dr. Anubha Shrivastava, "An efficient Software Source Code Metrics for Implementing for Software quality analysis", International Journal of Emerging Trends in Engineering Research, Volume 7, No. 9 September 2019.