Liver Patient Detection Using Machine Learning

Princy Prakash

Department of Computer Applications

Amal Jyothi College of Engineering,Koovapally

Kottayam, Kerala

Princyprakash2022@mca.ajce.in

Mr. BinumonJoesph

Department of Computer Applications

Amal Jyothi College of Engineering,Koovapally

Kottayam, Kerala

binumonjoesph@amaljyothi.ac.in

***Abstract* -Diagnosis of liver disease at early stage is vital for efficient therapy. It is a demanding issue in medical field to predict the disease in the preliminary stages owing to precise symptoms. Often the symptoms become apparent when it is too late. To get over this problem, this paper aims to improve this disease detection using some learning approaches. The objective is to research using classification algorithms to identify the liver patients from healthy personnel. This paper also focuses to conclude the classification algorithms based on its performance analysis.**

***Keywords*- machine learning, classification, visualization, correlation.**

I. INTRODUCTION

Issue with liver patients is not easily discoverable in an early stage because it will function normally even when it is partially damaged. An early diagnosis of liver problems can increase client's survival rate. Liver failures are at high rate of risk among most of the person. It is reported that by 2024 India will become the World Capital for Liver Diseases. The widespread occurrence of liver infection in India is contributed due to deskbound lifestyles, increased alcohol consumption and smoking. There are about 105 types of liver infections in this world. Therefore, developing a system that can improve the diagnosis of the malady will be a great milestone in the medical science. These systems will help the physicians in making well aimed decisions on patients and also with the help of reflex Classification tools for liver diseases one can definitely reduce the patient queues at the liver experts such as endocrinologists.

II. MACHINE LEARNING ALGORITHMS

Machine learning is the science of getting computers to behave without being programmed. It is a bough of synthetic intelligence and laptop technology which specializes in using facts and algorithms to act the manner that people research, step by step enhancing its accuracy. It is a technique this is used to analyze data that automates analytical model building. We implement Machine Learning based on the concept that the system could study from a given set of data, identify them and make decision with a low human interrelation. It focuses on data and algorithms to follow with way which humans learn and try and improve the accuracy. With the furtherance brought in Machine Learning, detection and prediction of liver cancer can be done with low effort. It can be used to diagnose not onlycancerous diseases but also other medical conditions. Machine Learning may be classified into 3 categories:

a. *Supervised Learning*
b. *Unsupervised Learning*
c. *Reinforcement Learning*

DOI: 10.5281/zenodo.6362045
ISBN: 978-93-5607-317-3 @2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam

Supervised Learning is the most basic among the other types. Here a model is trained on labeled data. It very extremely powerful enough to use in right circumstances. It will always continue to improve even if it is deployed. In unsupervised learning, we can learn the models even without labeling the data. Since there is no labeling it results hidden structures which makes this learning distinct from others. Reinforcement learning is different from the other two.

III . CLASSIFICATION

It is a technique of dividing given set of statistics to lessons, It may be completed on each established or unstructured statistics. The technique begins off evolved with forecasting the elegance of statistics points. The lessons are regularly known as labels. The type predictive modeling is the challenge of approximating the characteristic enter variables to discrete variables. The fundamental intention is to be aware of which elegance/class the brand new statistics will fall into. • Lazy Learners – Lazy rookie sincerely shop the schooling statistics and wait till a trying out statistics appears. The type is completed the use of the maximum associated statistics within side the saved schooling statistics. They have extra predicting time in comparison to rookies. Eg – k-nearest neighbor, case-primarily on totally reasoning. • Eager Learners – Eager rookies assemble a type version primarily based totally at the given schooling statistics earlier than getting statistics for predictions. It ought to be capable of decide to a unmarried speculation to be able to paintings for the complete space. Because of this, it take quite a few time in schooling and much few time in prediction. Eg – Decision Tree, Naïve Bayes, Artificial Neural Networks..

Logistic Regression

It is a class set of rules in gadget getting to know that makes use of one or greater unbiased variables to decideanfinal results. The final results is measured with a dichotomous variable that meansit's going to have handiestfeasibleoutcomes.Thepurpose of this regression is to discover a better-becomingcourtingamong the structured variable and a hard and fast of unbiased variables. It is higher than different binary class algorithms because it explains the elementsmain to class. Logistic regression is speciallysupposed for class.

The important drawback of the this regression set of rules is it most effective works while the expected one is binary, it guesses that the facts is freed from lacking values and guesses that predictors are impartial of each other. Naive Bayes Classifier is a class set of rules primarily based totally on Bayes's theorem which offers an assumption of independence amongst predictors. In easy terms, a Naive Bayes classifier assumes that the presence of a specific function in a category is unrelated to the presence of another function. Even though the capabilities rely on every other, all of those houses make a contribution to the opportunity independently. Naive Bayes version is simple to generate and is mainly beneficial for relatively massive facts. Even with a easiest approach, it is thought to outperform maximum of the class techniques in gadget learning. Following is the Bayes theorem to put in force the Naive Bayes Theorem.

Decision tree

The choice tree set of rules builds the typeversionwithinside theshape of a tree shape. It makes use of the if-then policieswhich canbesimilarly exhaustive and togetherdifferent in type. The techniqueis going on with partitioning the facts into simpler systems and subsequently joining it with an choice tree. The very lastshapeseems like a tree with nodes and leaves. The policies are found out sequentially the use of the educationfacts one at a time. Each time a rule is found out, the tuples masking the policies are removed. The techniquemaintainsat theeducation set till the termination factor is met.The tree is built in a top- down recursive divide and triumph over approach. A choice node can have or greater branches and a leaf represents a type or choice. The topmost node withinside thechoice tree that corresponds to the first-rate predictor is known asthe basis node, and the first- rateissueapproximately a choice tree is that it couldcope witheachexpress and numerical facts. A choice tree offersa bonus of simplicity to recognize and visualize, it calls forlittle or nofactsinstruction as well. The downside that follows with the choice tree is that it could create complicatedtimberwhich could bot categorize efficiently. They may beprettyriskydueto the fact even a simplistic extradewithinside thefacts can avert the entireshape of the choice tree.

Random Forest

Random choicebushes or random woodland are an ensemble getting to knowapproach for class, regression, etc.It operates via way of means ofbuildinga large number of choicebushes at schooling time and outputs the magnificencethis is the mode of the training or class or suggest prediction(regression) of the characterbushes. A random woodland is a meta-estimator that suitssome ofbushes on diverse subsamples of informationunitsafter whichmakes use ofa meanto enhance the accuracy withinside the model's predictive nature. The sub- patternlength is continuallysimilar to that of the authenticenterlengthhowever the samples are frequently drawn with replacements..

Support Vector Machine

The aid device is a classifier that depicts the schoolinginformation as factors in area divided into classesthroughan opening as extensive as possible. New factors are then brought to areathrough predicting which class they fall into and which areathey'll belong to.

Classifier Evaluation

The maximumvitalcomponent after the finishing touch of any classifier is the assessmentto test its accuracy and efficiency. There are a variety ofmethodswhereinwe willexamine a classifier. Let us checkthosestrategiesindexed below.

Holdout Method

This is the maximumnot unusual placetechniqueto assess a classifier. In this technique, the given statistics set is split into components as a take a look at and teach set 20% and 80% respectively.The teach set is used to teach the statisticsand the unseen take a look at set is used to check its predictive power.

Cross-Validation

In a cross validation technique a set of samples are divided into equivalent n subsamples. They are tested separately and study individually. The result then is combined togetherand form the final outcome.

IV .VISUALIZATION

Data visualization is described as a graphical illustration that includes the statistics and the statistics. By the usage ofvisiblefactors like charts.

V. CORRELATION

Correlation Matrix is largely a covariance matrix. Also referred to variance-covariance matrix. It is a matrix whereini-j function defines the correlation among the ith and jth parameter of the given records-set. When the recordsfactorsobserve a kind ofinstantly-line trend, the variables are stated tohave an about linear relationship. In a few cases, the recordsfactors fall nearainstantly line, howevergreaterfrequentlythere'sprettya chunk of variability of the factorsacross theinstantly-line trend. A precisdegreeknown as the correlation describes the power of the linear affiliation. Correlation summarizes the power and route of the linear (instantly-line) affiliationamong quantitative variables. Denoted with the aid of using r, it takes values among -1 and +1. A wonderfulcost for r shows a wonderfulaffiliation, and a poorcost for r shows a pooraffiliation. The nearer r is to one the nearer the recordsfactors fall to ainstantly line, thus, the linear affiliation is stronger. The nearer ris to 0, making the linear affiliation weaker.

VI . PROPOSED METHODOLOGY

The motive of this paper is to attain in a end to discover a Machine Learning versionthat mayexpectwhether or not the affected person have liver disaese or not.

a. DataSet

The dataset has been downloaded from the website Kaggle. The steps that are going to perform are: 1) Data Analysis: This is in standardsearchingon theinformation to determine out whats going on. Inspect the information: Check whether or notthere may be any lackinginformation, beside the pointinformation and do a cleanup. 2) Data Visualization: 3) Feature selection. 4) Search for any trends, relations & correlations. 5) Draw an inference and expectwhether or not the affected personmay bediagnosed to be having liver ailment or not

Table : dataset

Sl No.	Attributes
1	age
2	Total_bilirubin
3	Direct_bilirubin
4	Alkaline_phosphotase
5	Alamine_Aminotransferase
6	Aspartate_Aminotransferase
7	Total_Protiens
8	Albumin
9	Albumin_and_Globulin_Ratio
10	Dataset

The given data set includes 416 liver client details and 167 non liver client details collected from Andhra Pradesh. The Dataset column is a class label used to divide groups into liver patient and non liver patient.

Data Visualization

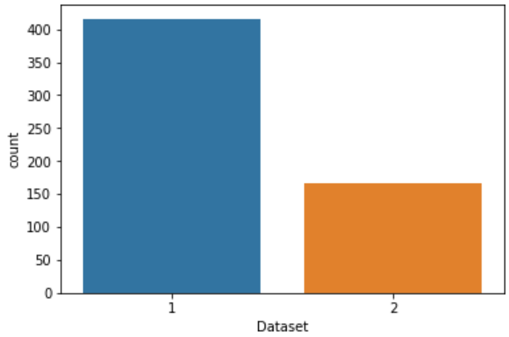


Figure 1:data visualization

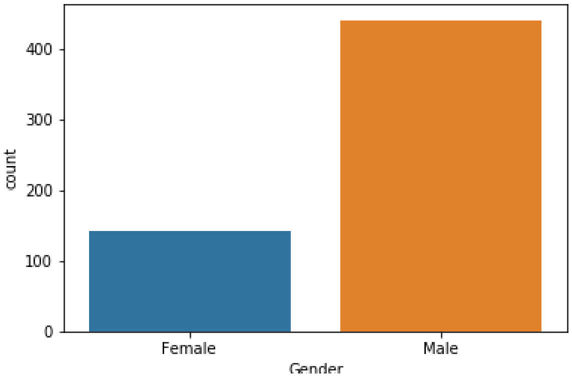


Figure 2: data Visualization

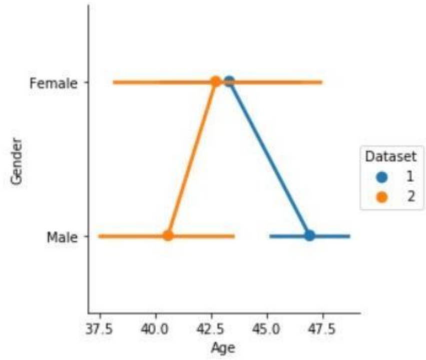


Figure 3 : factorplot

We can observe that age seems to be a major factor for the disease for both the genders

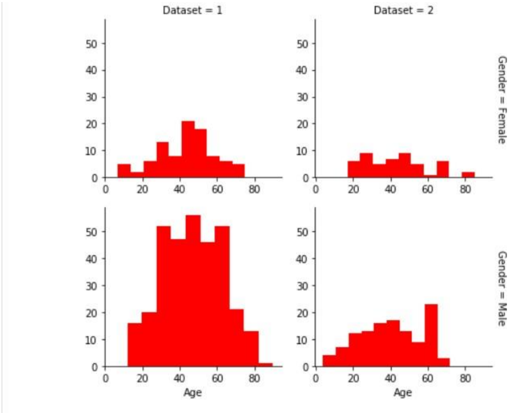


Figure 4 : Disease by gender and age

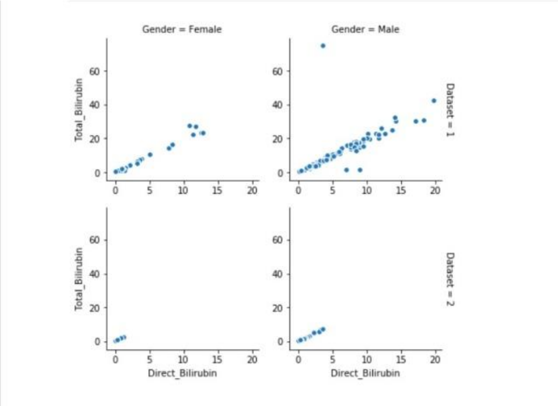


Figure 5 : scatterplot

There seems to be direct relation between some data. We can possibly remove one of this feature.

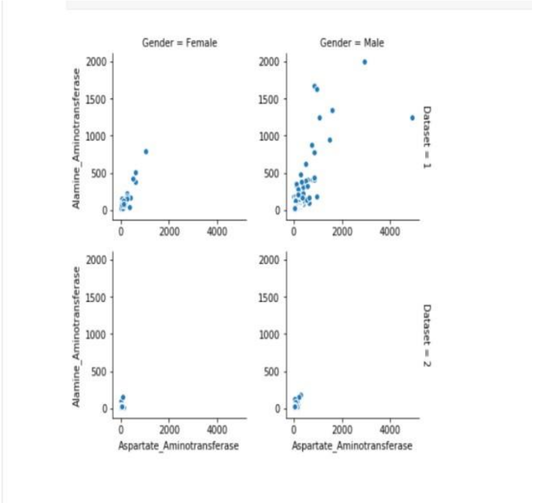


Figure 6 : scatterplot2

We can see a linear relation between some data and the gender. We can possibly remove one of this feature.

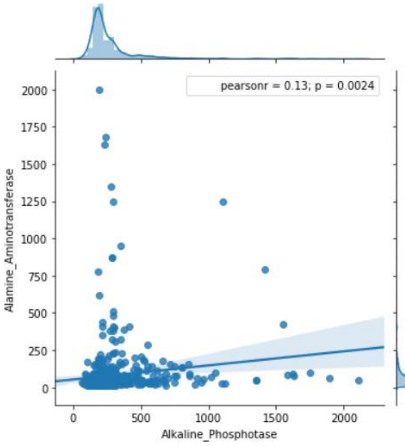
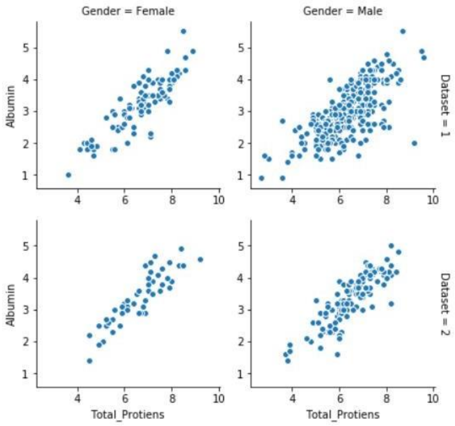


Figure 7: linear correlation

No linear correlation between Alkaline_Phosphotase and Alamine_Aminotransferase.



Observation:

From the above jointplots and scatterplots, we find direct relationship between the following features:Direct_Bilirubin & Total_BilirubinAspartate_Aminotransferase & Alamine_AminotransferaseTotal_Protiens & AlbuminAlbumin_and_Globulin_Ratio & Albumin Hence, we can very well find that we can omit one of the features. I'm going to keep the follwing features:Total_BilirubinAlamine_AminotransferaseTotal_ProtiensAlbumin_and_Globulin_RatioAlbumin

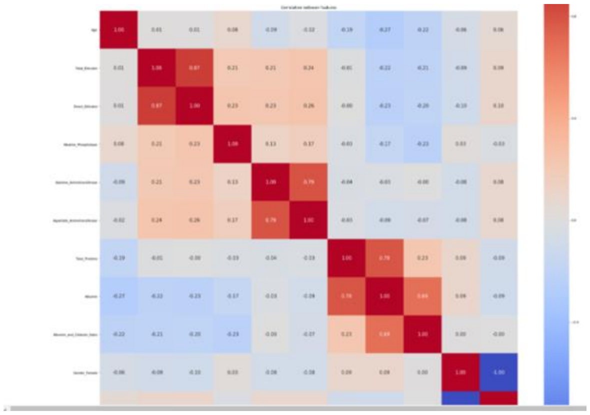


Figure 9 : correlation

DOI: 10.5281/zenodo.6362045
ISBN: 978-93-5607-317-3 @2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam

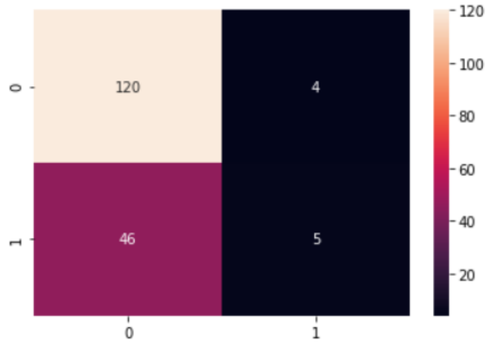


Figure 10:Confusion matrix

VII. CONCLUSION

Health issue related to liver is becoming more common nowadays. With repeated technological enhancements, these maladies can be reduced in future. Even though people are becoming pretty conscious about their health nowadays, the uncontrolled lifestyle and luxuries that are incrementally being emerged and enhanced, the problem is going to last long. In the end we get a confusion matrix that gives a total accuracy of 71.42%.

VIII. REFERENCES

[1] Software based prediction of liver disease with feature selection and classification techniques written Jagdeep Singh, Sandeep Bagga and Ranjodh Kaur.

[2] Prediction and analysis of liver diseases using data mining written Shambel Kefelgen, Pooja Kamat.

[3] Strategic analysis in prediction of liver disease using classification algorithms written by Piyush Kr Shukla andBinish Khan

[4] https://www.kaggle.com/sanjames/liver-patients-analysis-prediction-accuracy/notebook

[5] https://www.ijert.org/liver-disease- prediction-system-using-machine-learning-

DOI: 10.5281/zenodo.6362045
ISBN: 978-93-5607-317-3 @2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam