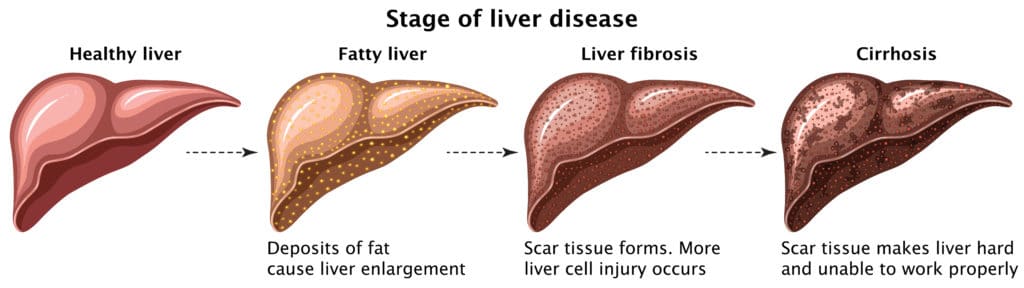
**A close-up of a logo

Description automatically generated with medium confidence**

**Liver cirrhosis prediction using XGboost and Stratified K-fold**



|  |  |
| --- | --- |
| **Student**  **Vinay Ale**  **0784374** | **Instructor**  **Nicole Sbrocca**  **Krista Naccarato** |

**Contents:**

* **Title**
* **Introduction**
* **Related work**

1. Prediction of liver cirrhosis, using diagnostic imaging tools
2. Important predictor of mortality in patients with end-stage liver disease

* **Data description**
* **Methods**

1. Exploratory Data Analysis
2. Handling missing values
3. Insights
4. Model building

* **Evaluation Metrics**

1. Accuracy
2. Precision
3. Recall
4. F1-Scores

* **Results:**

1. Best Accuracy

* **Discussion:**

1. Stratified K-Fold
2. Logistic Regression
3. XGBoost

* **Conclusion**
* **Appendix**
* **References**

**Introduction:**

Healthcare Analytics is a process of analysing current and historical industrial data to predict trends, improve outreach and even better analyse the disease to manage the spread of disease.

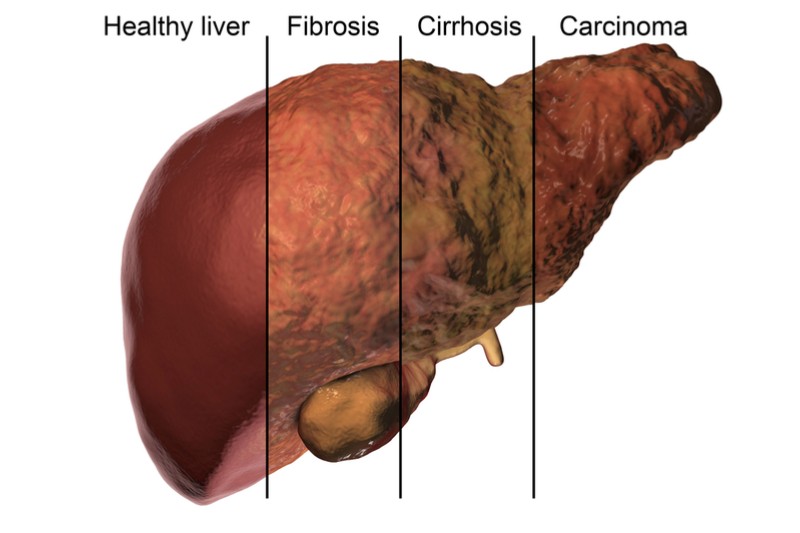
Liver cirrhosis is a chronic disease in which there is a continuous long-term damage of liver where, the normal cell of liver is replaces with scar tissues. The abnormal tissues like scab that forms superficial skin wound, but much tougher and stronger. As the normal cells are replaced with chaff (scar tissue) liver will gradually losses its functions and overtime fails.

Cirrhosis is caused by wide range of diseases and conditions. As we know one of the main reasons is alcohol abuse, viral hepatitis, fat accumulation in liver, genetic digestive disorder, moreover immune system also causes liver cirrhosis and so on.

There are some symptoms to identify early like, Fatigue, easily bleeding, Loss of appetite, Nausea, swelling in legs, feet and ankles, itching skin, weight loss and jaundice and many other.

In this project we developed an algorithm where it will predict the liver cirrhosis in the early stage so that it will be helpful for physicians to understand what the stage of cirrhosis is. Such that, what kind of treatment should be given to patient.

**Stages of cirrhosis**



**Related works:**

**Paper1:**

**Prediction of liver cirrhosis, using diagnostic imaging tools**

It's crucial to diagnose cirrhosis of the liver early. The gold standard for cirrhosis of the liver diagnosis is ultrasonic liver biopsy. However, the method's applicability is constrained by its intrusiveness and sample bias. Over the past few decades, basic imaging for the diagnosis of liver cirrhosis has advanced, making it possible to detect early morphological abnormalities in the liver using ultrasonography (US), computed tomography, and magnetic resonance imaging (MRI) (MRI). They are also reliable diagnostic techniques for advanced liver cirrhosis, for which it is challenging to make an early diagnosis. The homogeneity of the hepatic parenchyma can be more precisely identified through texture analysis, elastography can be used to quantify the stiffness and elasticity of the liver, and perfusion studies can be used to calculate the volume, transit time, and velocity of the blood flow. Amongst these methods, elastography using US and MRI was found to be slightly easier, faster, and able to provide an accurate diagnosis. Early diagnosis of liver cirrhosis using MRI or US elastography is therefore a realistic alternative, but further research is still needed.

**Paper2:**

# Important predictor of mortality in patients with end-stage liver disease

Any disease's baseline assessment must include the prognosis. Numerous prognostic models have been presented for predicting the prognosis of end-stage liver disease. For over three decades in the context of end-stage liver disease, the Child-Pugh score has served as the standard for determining the prognosis of cirrhosis. Despite several restrictions, a recent large systematic review demonstrated that Child-Pugh score and its components (bilirubin, albumin, and prothrombin time) were still reliable predictors. Child-Pugh score was recently replaced as a "contemporary" option by Model for end-stage liver disease (MELD) score. The MELD score has proved crucial for determining the severity of liver disease and for assessing the mortality risk. MELD score has a few weaknesses, hence new, modified MELD scores (MELD-Na, Delta MELD) have been developed and validated. This review summarizes the current knowledge about the prognostic factors in end-stage liver disease, focusing on the role of Child-Pugh and MELD score.

**Data Description:**

The data is collected from the mayo clinic trails in primary biliary cirrhosis (PBC) of the liver. These studies are conducted in the years 1974 and 1984. 424 PBC patients who were sent to Mayo Clinic during that ten-year period qualified for the drug D-penicillamine's placebo-controlled trial.

The dataset consists of following columns :

1. ID: unique identifier
2. N\_Days: number of days between registration and the earlier of death, transplantation, or study analysis time in July 1986
3. Status: status of the patient C (censored), CL (censored due to liver tx), or D (death)
4. Drug: type of drug D-penicillamine or placebo
5. Age: age in [days]
6. Sex: M (male) or F (female)
7. Ascites: presence of ascites N (No) or Y (Yes)
8. Hepatomegaly: presence of hepatomegaly N (No) or Y (Yes)
9. Spiders: presence of spiders N (No) or Y (Yes)
10. Edema: presence of edema N (no edema and no diuretic therapy for edema), S (edema present without diuretics, or edema resolved by diuretics), or Y (edema despite diuretic therapy)
11. Bilirubin: serum bilirubin in [mg/dl]
12. Cholesterol: serum cholesterol in [mg/dl]
13. Albumin: albumin in [gm/dl]
14. Copper: urine copper in [ug/day]
15. Alk\_Phos: alkaline phosphatase in [U/liter]
16. SGOT: SGOT in [U/ml]
17. Triglycerides: triglicerides in [mg/dl]
18. Platelets: platelets per cubic [ml/1000]
19. Prothrombin: prothrombin time in seconds [s]
20. Stage: histologic stage of disease (1, 2, 3, or 4)

**Methods:**

**Exploratory Data Analysis:**

The dataset can be described in the approaches section. Using summary statistics and graphical representations, exploratory data analysis is a crucial method for identifying patterns, identifying anomalies, testing theories, and validating assumptions. Following the import of the necessary libraries, including NumPy, we loaded the data set using pandas. There is a count, mean, minimum and maximum value, and standard deviation.

To understand the dataset and to displays few values from dataset multiple functions are used. Such as,

* Head(): this is used to display the no of rows want to display.
* Info() : is used to display no of non-null values in each column. So that it will be helpful for finding missing values in each column.
* **Describe():**

This function is used to get a descriptive statistics summary of a data frame.

This includes the percentiles, min, max, mean, median and mode of each variable.

* **Isna() and Sum():**

Isna() functions is used to find null values in each variable and return as Booleans and sum() function is used to sum the no of null values in variable. Together it returns no of null values.

**Handling missing values:**

This is a problem, we could just get rid of all examples with NA values, but in this case our case of small dataset we cannot afford that.

We will impute the missing entries with some statistical calculations. Such as, meadian for numerical variables.

**We have two different types of data**

1. Numerical data (Age, Cholesterol, Platelets. etc)
2. Categorical Data (Drug, Sex, Spiders. Etc)

We will have to use different imputation for each type

1. For the numerical type we can use mean or median. In this case we will go with median to avoid skewing in the presence of outliers
2. For Categorical type we will impute the most frequent class.

**Insights:**

As we can observe we have class imbalances in our dataset i.e some classes have more examples than other. This could make it difficult for our model to train and achieve desired score. No worries, we can fix that later.

Icon

Description automatically generated

There are 4 stages of cirrhosis in dataset which is our target variable. Where most of the patients are in stage 3 and 4.

**Graphical user interface, application

Description automatically generated**

The above graphs show the relationship between the target variable stage and the other variables like Drug, status, sex, ascites, hepatomegaly, Spiders and Edema.it can be seen that cirrhosis is mostly seen in males compare to females.

**Model Building:**

In this section we try to set the parameters into feature variables as (X) and target variable as (y). Here 2 machine learning algorithms are used XGboost and logistic regression with stratified K-fold. different models are used to check for the better accuracy.

**Evaluation Metrics**:

Metrics for evaluation are used to measure how well a statistical or machine learning model is performing. Every project has to evaluate machine learning models or algorithms. To test a model, a wide variety of evaluation measures are available.

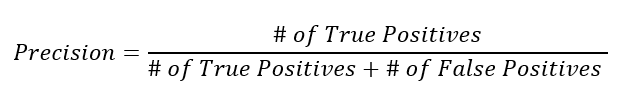
Accuracy is often the most widely used statistic, however because our project's data is so severely out of balance, we were able to achieve superior outcomes with F1 score.

**Accuracy:**

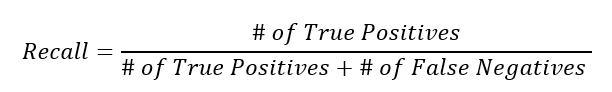
[Text

Description automatically generated](https://towardsdatascience.com/the-f1-score-bec2bbc38aa6)Accuracy is defined as number of correct predictions to the total no of predictions. Accuracy is one of the mostly used metrics for machine learning classification models.

**Precision**:

[](https://towardsdatascience.com/the-f1-score-bec2bbc38aa6)Precision is a metrics which shows the accuracy of positive predictions. Precision is calculated as the ratio of true positives to all positive predictions.

**Recall:**

[](https://towardsdatascience.com/the-f1-score-bec2bbc38aa6)Recall is a metric used to measure how many correct positive predictions were produced out of all possible positive predictions.

**F1-Scores:**

[Text, letter

Description automatically generated](https://towardsdatascience.com/the-f1-score-bec2bbc38aa6)F1-score is defined as harmonic mean of precision and recall. It is which is used for statistical measurements.

**Results:**

|  |  |  |
| --- | --- | --- |
| **Model** | **Accuracy** | **F1 Score** |
| XGboost | 78% | 84% |
| Logistic regression | 70% | 74% |

**Best Accuracy:**

The Best accuracy was achieved with XGboost.

Table

Description automatically generated

Here, the accuracy is 78% using the XGboost. So that, using XGboost model we predict cirrhosis with 78% accurate.

**Discussion:**

**Stratified k-fold:**

Stratified K-Fold cross validator is used to provide the train and test values to split the data in train and test sets for K times. The cross-validator object is a K-Fold which returns stratified folds. The folds are made the percentage of samples in each class. This will generate the test set by smallest and largest differ the most sample. This is better than random sampling because each time the dataset is split into train and test data. Moreover, provides the values to the model for prediction.

Text

Description automatically generated**Logistic Regression:**

Based on one or more predictor variables, logistic regression is used to predict the class (or category) of individuals (x). It is used to simulate a binary result, or a variable with only two potential values, such as 0 or 1, yes or no, or diseased or not.

Text

Description automatically generated

Table

Description automatically generatedHere, using logistic regression with stratified K-fold there are different accuracies the average of all accuracies is 70%.

The precision score is 70% for predicting 0’s and 45% for predicting 1’s.

The Recall Score is 78% for predicting 0’s and 36% for 1’s.

The f1-score is 74% for 0’s and 40% for 1’s.

**XGboost classifier:**

Text

Description automatically generatedXGboost is an implementation of Gradient Boosted Decision trees. Decision trees are sequentially created. Weights plays major role in XGBoost. All independent variables are assigned with weights which are fed into decision tree which predicts results.

Based on one or more predictor variables, XGBoost is used to predict the class (or category) of individuals (x). It is used to simulate a binary result, or a variable with only two potential values, such as 0 or 1, yes or no, or diseased or not.

**Table

Description automatically generated**Here, using logistic regression with stratified K-fold there are different accuracies the average of all accuracies is 78%.

The precision score is 82% for predicting 0’s and 69% for predicting 1’s.

The Recall Score is 85% for predicting 0’s and 64% for 1’s.

The f1-score is 84% for predicting 0’s and 67% for predicting 1’s.

**Conclusion:**

The process of analysing the historical data related to patients and hospitals are known as Healthcare analysis. In this project the data that is collected from Mayo Clinic trail in primary biliary cirrhosis (PBC) of liver which is conducted in the year 1974 to 1984. Cirrhosis is not a contagious disease but it is dangerous once effect. There are many symptoms to recognise the disease but, Using machine learning models we can accurately predict the disease in the starting stage accurately. XGBoost and Logistic Regression models are used for predictions. The best result achieved By XGBoost which is 73%. Which means cirrhosis is predicted using XGboost by 73% accurately.

**Appendix:**

**References**

Images:

<https://www.medicdrive.org/liver-cirrhosis/>

<https://www.labiotech.eu/more-news/versantis-liver-cirrhosis-treatment/>

Introduction

<https://www.mayoclinic.org/diseases-conditions/cirrhosis/symptoms-causes/syc-20351487>

<https://www.medicdrive.org/liver-cirrhosis/>

Related works

Paper1: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4539400/>

Yeom SK, Lee CH, Cha SH, Park CM. Prediction of liver cirrhosis, using diagnostic imaging tools. World J Hepatol. 2015 Aug 18;7(17):2069-79. doi: 10.4254/wjh.v7.i17.2069. PMID: 26301049; PMCID: PMC4539400.

Paper2: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3701842/>

Kim HJ, Lee HW. Important predictor of mortality in patients with end-stage liver disease. Clin Mol Hepatol. 2013 Jun;19(2):105-15. doi: 10.3350/cmh.2013.19.2.105. Epub 2013 Jun 27. PMID: 23837134; PMCID: PMC3701842.

Eavaluation Metrics:

<https://medium.com/analytics-vidhya/how-to-evaluate-your-machine-learning-model-76a7671e9f2e>

XGboost model:

<https://www.geeksforgeeks.org/xgboost/?ref=gcse>