# Using machine learning methods to diagnose Hepatitis C infection

## 1. Abstract:

In this project, I aim to make diagnostic decision whether a person has Hepatitis C or he/she is just a blood donor. There are three stages of Hepatis C in which a person can be classified further. These represent the progress of a person into Hepatitis C. The initial stage is tagged with just 'Hepatitis', mild stage is 'Fibrosis' and 'Cirrhosis' is the chronic hepatitis C infection. A blood donor can also be classified further into a blood donor or a suspect blood donor. I will use random forests and decision trees machine learning algorithm to attempt to make a diagnosis of Hepatitis C. The predictions made have good accuracy.

## 2. Introduction

### 2.1 Contextualization: Hepatitis C

Hepatitis C is a virus that can infect the liver. This virus can cause both acute and chronic hepatitis. If left untreated, it can sometimes cause serious and potentially life-threatening damage to the liver over many years. Hepatitis is a major cause of liver cancer.

Hepatitis C virus is a bloodborne virus and a healthy person can get infected by coming in contact with the blood of an infected person. This may happen through injection drug use, unsafe injection practices, unsafe health care, transfusion of unscreened blood and blood products, and sexual practices that lead to exposure to blood.

Antiviral medicines can cure more than 95% of persons with hepatitis C infection, thereby reducing the risk of death from cirrhosis and liver cancer, but access to diagnosis and treatment is low. There is currently no effective vaccine against hepatitis C. [1]

### 2.2 Motivation: Diagnosis and prevention

Globally, an estimated 71 million people have chronic hepatitis C virus infection. A significant number of those who are chronically infected will develop cirrhosis or liver cancer. WHO estimated that in 2016, approximately 399,000 people died from hepatitis C, mostly from cirrhosis and liver cancer.

Following initial infection, approximately 80% of people remain asymptomatic. Because of this, only few people are diagnosed with hepatitis C when the infection is new. In those people who go on to develop chronic HCV infection, the infection is also often undiagnosed because it remains asymptomatic until decades after infection when symptoms develop secondary to serious liver damage. So regular monitoring for early diagnosis can prove to be a great factor in reducing the hepatitis C infection and preventing the transmission of the virus.

There are specific tests like anti-HCV antibodies test and HCV ribonucleic acid (RNA) test for diagnosing HCV infections. Apart from this, there are a series of special blood tests that can often determine whether or not the liver is functioning properly. These can distinguish between wide range of liver infections.

These liver function tests help determine the health of liver by measuring the level of proteins, liver enzymes, bilirubin, etc in the blood. So, these are the common tests done to detect any liver problems. This project utilizes the data, which is based on these traditional biochemical tests to diagnose hepatitis C infection. The names of these tests are ALB (albumin), ALP (alkaline phosphates), ALT (alanine amino-transferase), AST (apartate amino-transferase), BIL  (bilirubin), CHE (choline esterase), CHOL (cholesterol), CREA (creatinine), GGT (γ-glutamyl-transferase), and PROT (protein).

Laboratory diagnostic pathways are based on expert rules (“ if ….then….else”) [2]. So, machine learning algorithms especially like decision trees can easily be employed to make the diagnosis of hepatitis C infection. Being able to make early diagnosis of hepatitis C infection, can help to treat this infection in an earlier stage and to reduce the loss of life.

### 2.3 Project Aim

The aim of this project is to use machine learning models to diagnose if and which stage of hepatitis C infection a person has. The project is based on the HCV dataset available on UCI. The data set contains laboratory values of blood donors and Hepatitis C patients and demographic values like age, sex etc. The target attribute for classification is Category (blood donors vs. Hepatitis C (including its progress ('just' Hepatitis C, Fibrosis, Cirrhosis). Decision trees and random forests are the two machine learning algorithms that are going to be used to achieve the aim of diagnosis.

## 3. Literature review

As described in Hoffmann and Bietenbeck’s paper, machine learning algorithms should definitely be tried to see how they can contribute to enhancing the currents available diagnostic ways. decision trees can mimic the expert rules (if…then…else classification) which clinical and laboratory experts use to diagnose hepatitis C and its various stages. It can save costs, simplify the process and minimize the number of false positive and false negative results. Decision trees can either be used in deducing new potential trees or in validating existing ones by human experts [2].

Literature review of chronic hepatitis C reveals that individual biochemical tests are good at classifying the cases with no or minimal fibrosis from those with severe fibrosis or cirrhosis, but they are poor at predicting intermediate levels of fibrosis [4]. Same was the conclusion for decision trees that they were also unable to perfectly separate the different stages of hepatitis. Also, the results from decision trees may somewhat look arbitrary because the algorithms do not consider the scientific plausibility [2]. However, owing to the benefits machine learning algorithms can offer, it is worth to research and experiment their application in the area. That is why in this project, I attempt to further improve the efficacy of Hepatitis C prediction by utilizing the machine learning methods and computation resources available to us.

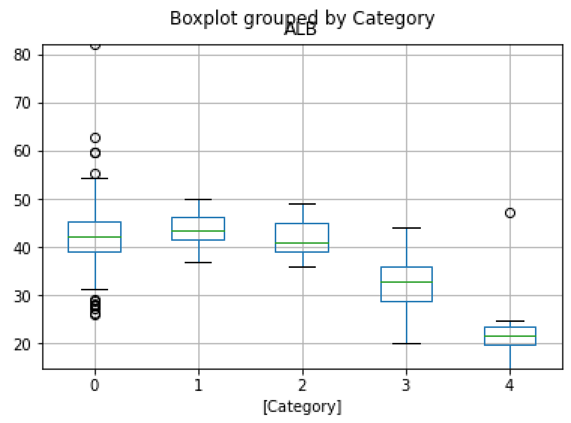
## 4. Methods

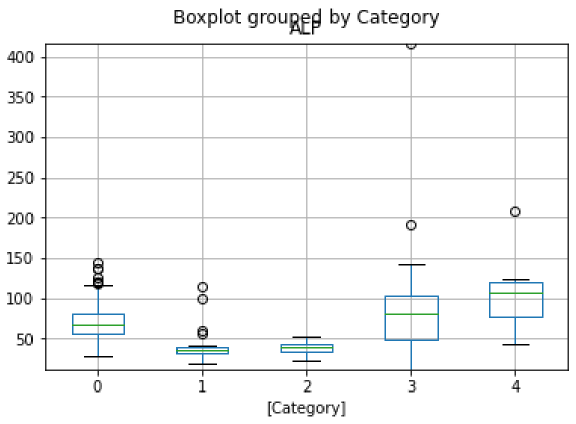
The aim of the project can be achieved by a machine learning classifier. From the literature review, it is clear that decision trees are a great choice to implement the diagnosis classifier for hepatitis C. This is because laboratory diagnostic pathways are based on expert rules (“ if ….then….else”), which is exactly a decision tree does at each node. They also offer an advantage over black box algorithms like support vector machines or neural networks that they can be easily evaluated by medical experts [2]. An ensemble of decision trees can offer great improvement in the prediction accuracy, hence random forests are also used in the project.

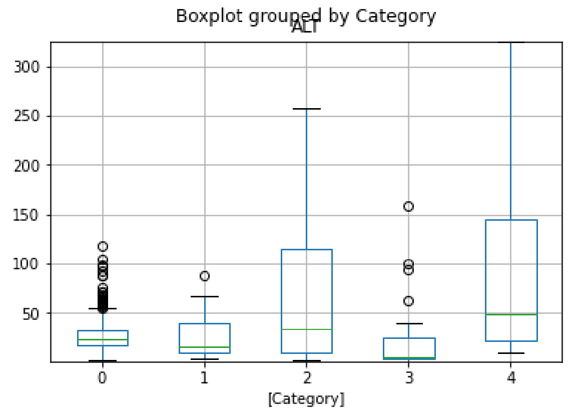
## 5. The Data

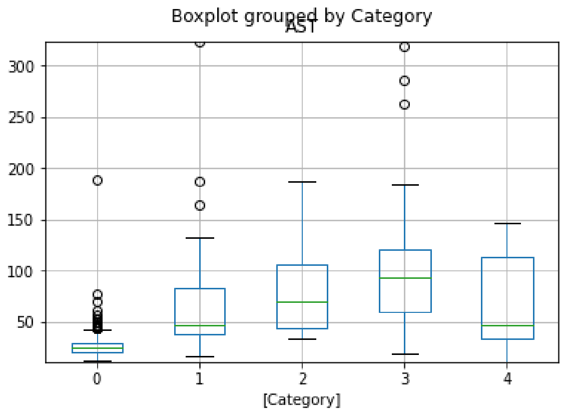
### 5.1 Exploratory data analysis

There are five categories of classification. '0=Blood Donor', '1=Hepatitis', '2=Fibrosis', '3=Cirrhosis', '4=suspect Blood Donor'. The dataset has the following attributes/features: Age (in years), Sex (female/male), ALB (albumin), ALP (alkaline phosphates), ALT (alanine amino-transferase), AST (apartate amino-transferase), BIL  (bilirubin), CHE (choline esterase), CHOL (cholesterol), CREA (creatinine), GGT (γ-glutamyl-transferase), PROT (protein). Some descriptive statistics using box plots can suggest the importance of various attributes in various categories of diagnosis. The following box plots gives the descriptive statistics of the available data:









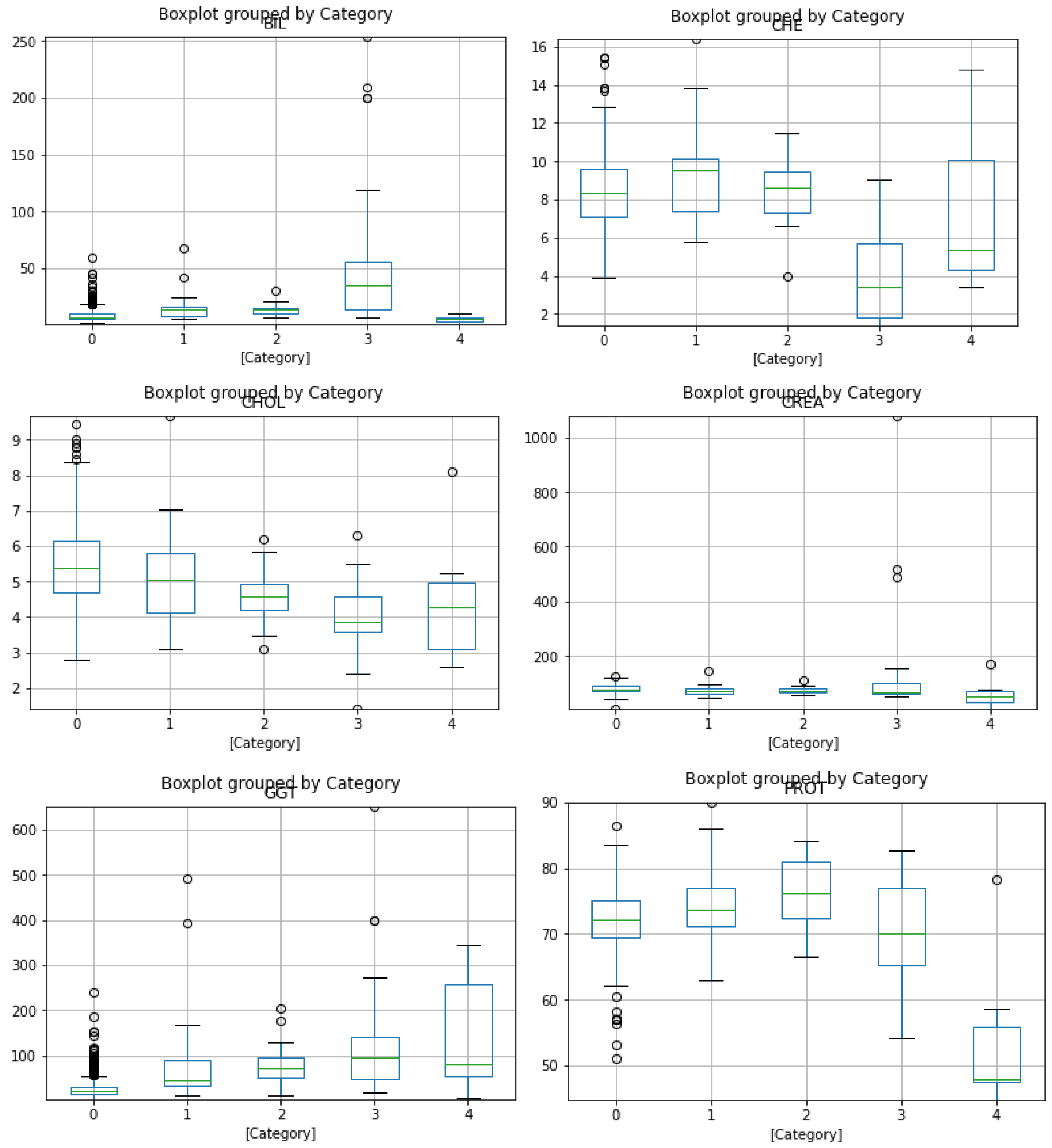
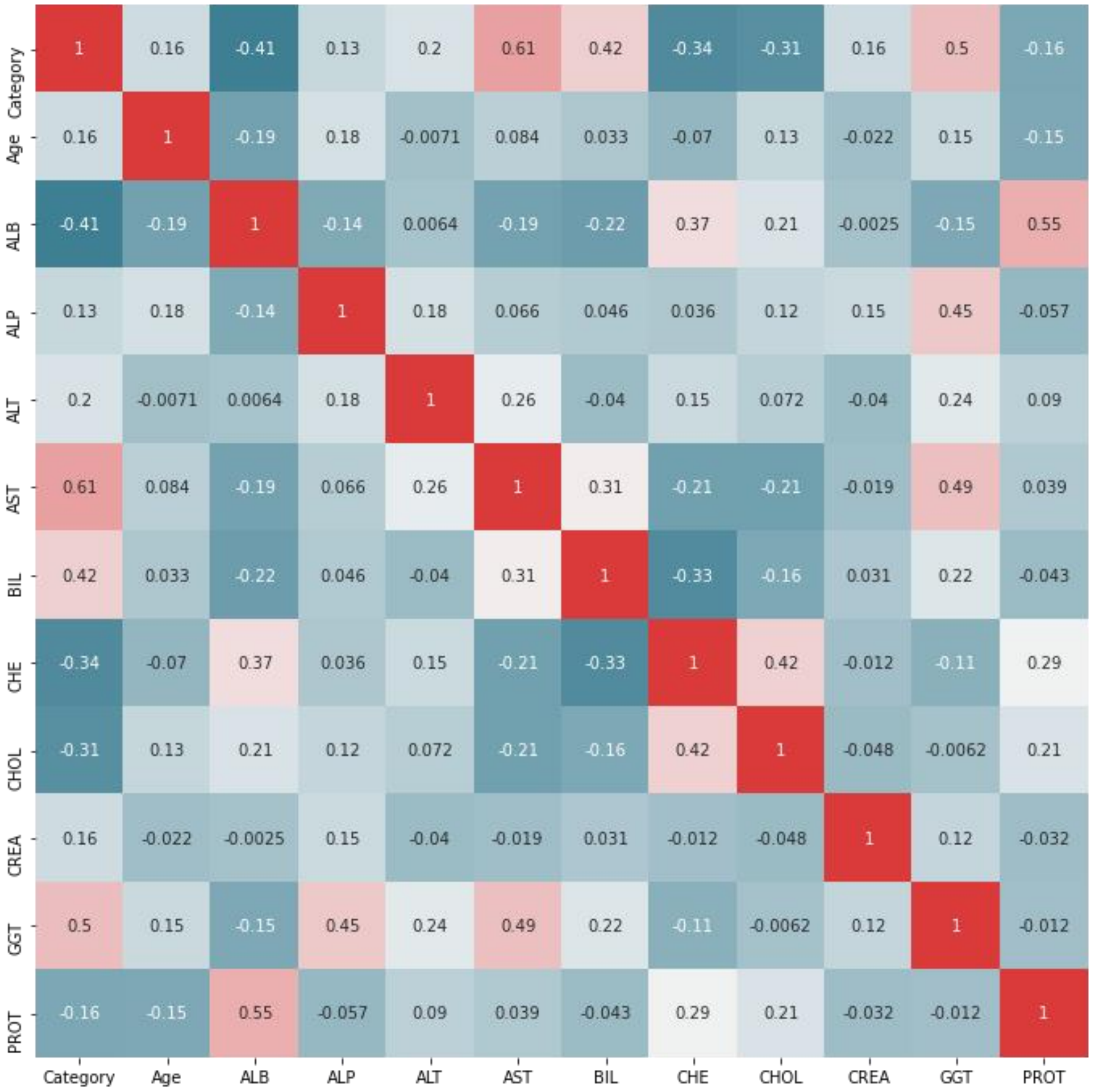


Fig. Descriptive statistics: Box plots of 10 attributes in each of the 5 classes

## Correlation among the attributes

The correlation matrix was plotted to get understanding of how various attributes of the data are correlated. The plot is as shown below.

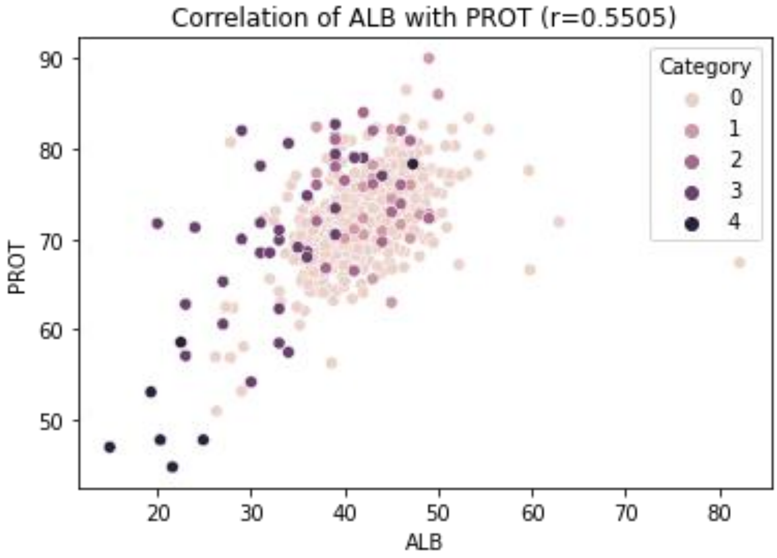


| | Category | Age | ALB | ALP | ALT | AST | BIL | CHE | CHOL | CREA | GGT | PROT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Category | 1 | 0.16 | -0.41 | 0.13 | 0.2 | 0.61 | 0.42 | -0.34 | -0.31 | 0.16 | 0.5 | -0.16 |
| Age | 0.16 | 1 | -0.19 | 0.18 | -0.0071 | 0.084 | 0.033 | -0.07 | 0.13 | -0.022 | 0.15 | -0.15 |
| ALB | -0.41 | -0.19 | 1 | -0.14 | 0.0064 | -0.19 | -0.22 | 0.37 | 0.21 | -0.0025 | -0.15 | 0.55 |
| ALP | 0.13 | 0.18 | -0.14 | 1 | 0.18 | 0.066 | 0.046 | 0.036 | 0.12 | 0.15 | 0.45 | -0.057 |
| ALT | 0.2 | -0.0071 | 0.0064 | 0.18 | 1 | 0.26 | -0.04 | 0.15 | 0.072 | -0.04 | 0.24 | 0.09 |
| AST | 0.61 | 0.084 | -0.19 | 0.066 | 0.26 | 1 | 0.31 | -0.21 | -0.21 | -0.019 | 0.49 | 0.039 |
| BIL | 0.42 | 0.033 | -0.22 | 0.046 | -0.04 | 0.31 | 1 | -0.33 | -0.16 | 0.031 | 0.22 | -0.043 |
| CHE | -0.34 | -0.07 | 0.37 | 0.036 | 0.15 | -0.21 | -0.33 | 1 | 0.42 | -0.012 | -0.11 | 0.29 |
| CHOL | -0.31 | 0.13 | 0.21 | 0.12 | 0.072 | -0.21 | -0.16 | 0.42 | 1 | -0.048 | -0.0062 | 0.21 |
| CREA | 0.16 | -0.022 | -0.0025 | 0.15 | -0.04 | -0.019 | 0.031 | -0.012 | -0.048 | 1 | 0.12 | -0.032 |
| GGT | 0.5 | 0.15 | -0.15 | 0.45 | 0.24 | 0.49 | 0.22 | -0.11 | -0.0062 | 0.12 | 1 | -0.012 |
| PROT | -0.16 | -0.15 | 0.55 | -0.057 | 0.09 | 0.039 | -0.043 | 0.29 | 0.21 | -0.032 | -0.012 | 1 |

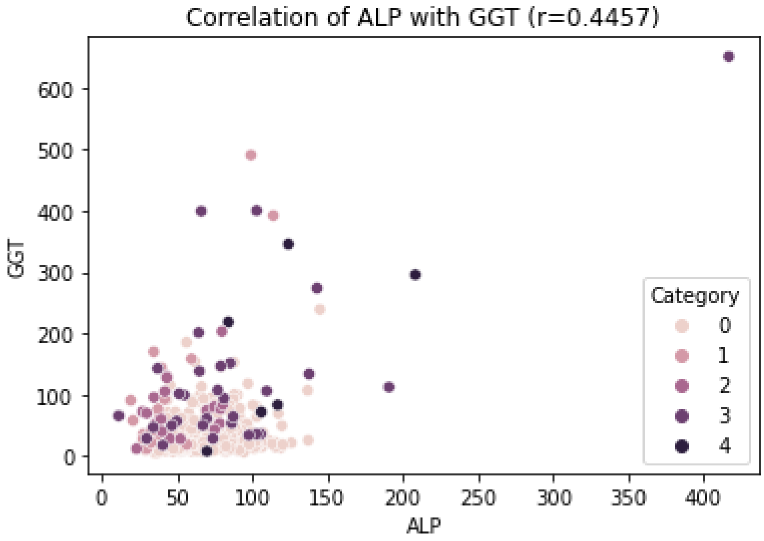
Fig. Plot of correlation matrix between various attributes of data

Observation:

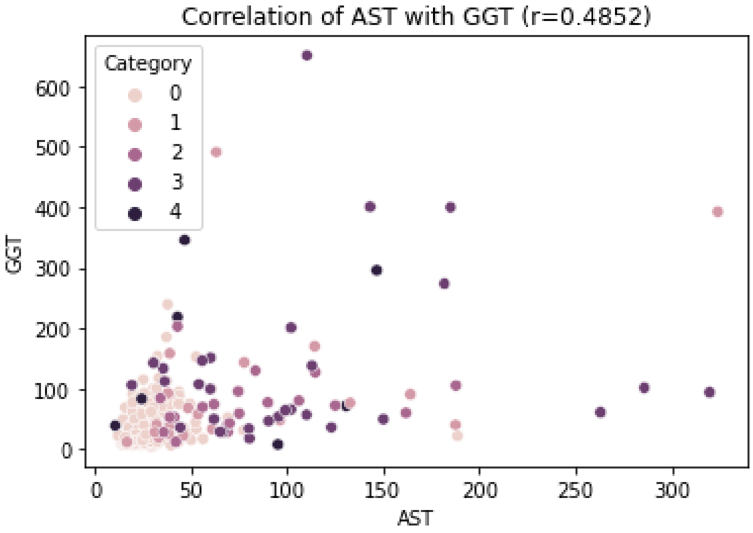
On studying the correlation coefficients, it is observed that following attributes are quite correlated. The scatter plots were drawn to further solidify the observation. One of these pairs is negatively correlated.
1) 'ALB’ and 'PROT'



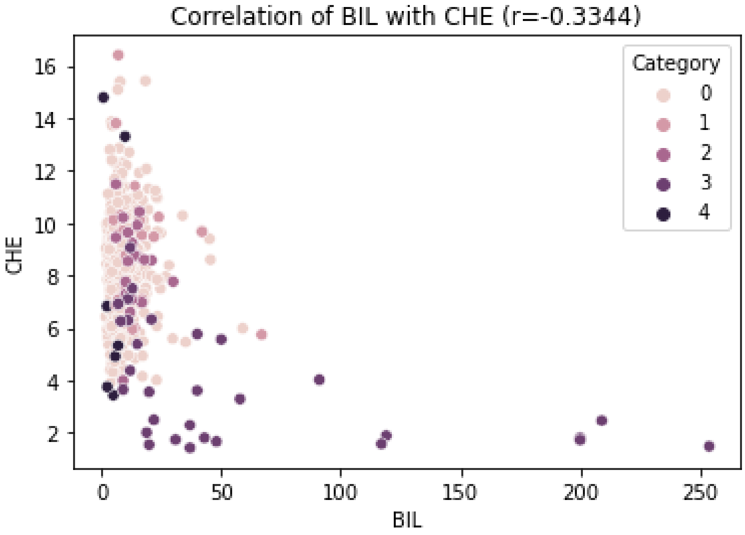
2) 'ALP' and 'GGT'



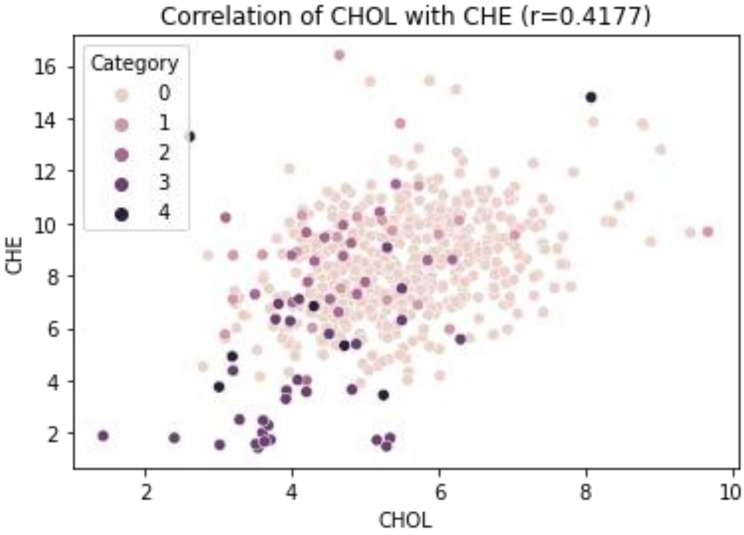
3) 'AST' and 'GGT'



4) 'BIL' and 'CHE'



5) 'CHOL' and 'CHE'



Unfortunately, there were some challenges related to the dataset. The data set is very small. It has only 615 data instances. It has acute class imbalance. It also has some missing data.

### 5.2 Missing Data

Some of the attributes like ALB, ALP, ALT, CHOL, and PROT has some missing data. In total, around 4.2 % of the entries have one or missing value. Most of the entries were missing in ALP and CHOL attributes.

### 5.3 Class imbalance

The data has class imbalance. There is a large number of instances for a particular class. This class is 'Blood donor'. So, accuracy is not a well-suited measure for evaluation of the model trained for this data set. Instead, precision, recall and fscore would be more suited in this case.

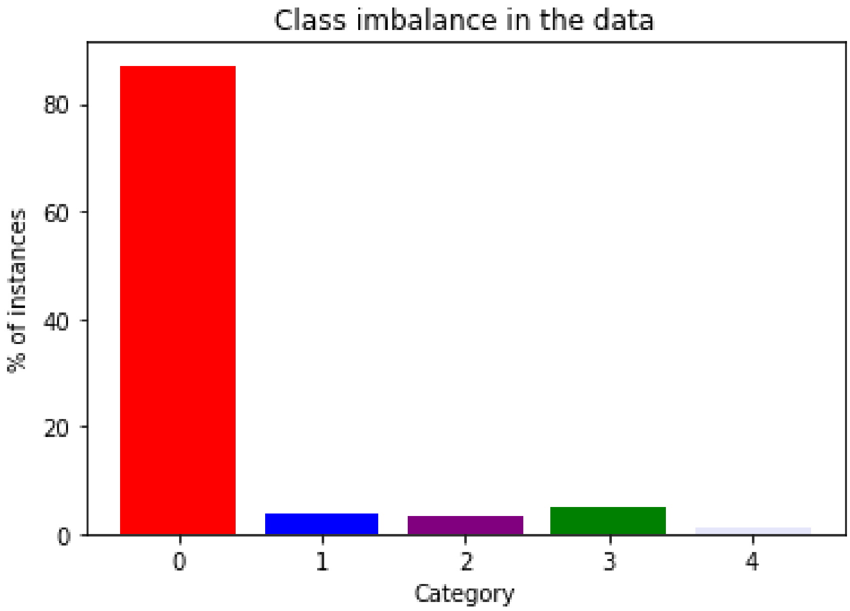


Fig. Plot depicting class imbalance

There are several methods of dealing with the class imbalance problem. Resampling the dataset is one of the main methods that can be used here. These are of two types:

- Under sampling
- Oversampling

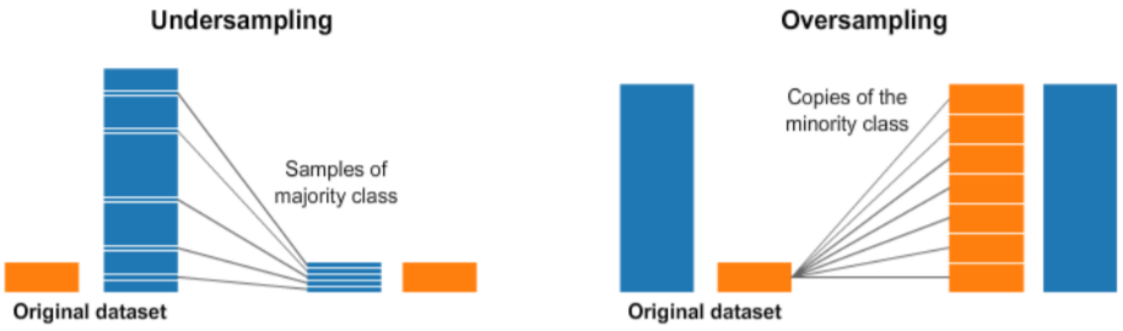


Fig. Explaining under sampling and over sampling methods [3]

In under sampling, the number of instances of the class with most representation is reduced such that it becomes comparable to the other classes. Whereas in over sampling, the samples of the classes which have lesser number of instances are increased such that the representation of all classes become almost equal.

In this project, experiments were performed with both methods to come up with a more accurate model.

### 5.4 Data pre-processing

The missing data was appropriately dealt with by replacing the missing values with the mean conditional on the groups formed by age and sex attributes. The data was also standardized before it was used for data analysis and machine learning.

### 5.5 Model selection

As the data set has only 12 attributes, it would be sufficient to use decision tree or random classifier here. Also, these classifiers are also good enough to deal with the multi-class classification. However, as stated before there are two major concerns. One is of the data set being small and the other of the class imbalance with a particular class being the most represented. To find out the configuration of a best performing classifier, many experiments were conducting. Some of them dealt with class imbalance by oversampling or undersampling, while others were specific to tuning the performance of the classifier. F score is used to measure the accuracy performance of a model for the reason already discussed above. In the end, the performances of various classifiers are compared.

## 6. Experimentation results

Various experiments were conducted to achieve good accuracy of the model. These are described below.

<u>Base Decision tree model without handling class imbalance</u>

The available data set has class imbalance. Mostly, there are blood donors among all the patients, hence Blood Donor class is the most represented class in this data set. Just to establish a baseline decision tree classifier was trained and evaluated on the data set directly, without handling the class imbalance. It gave the F-score of 90.2%, which seems like a good accuracy. However, it is mainly because of the underlying class imbalance, which is clearly depicted in the confusion matrix shown below.

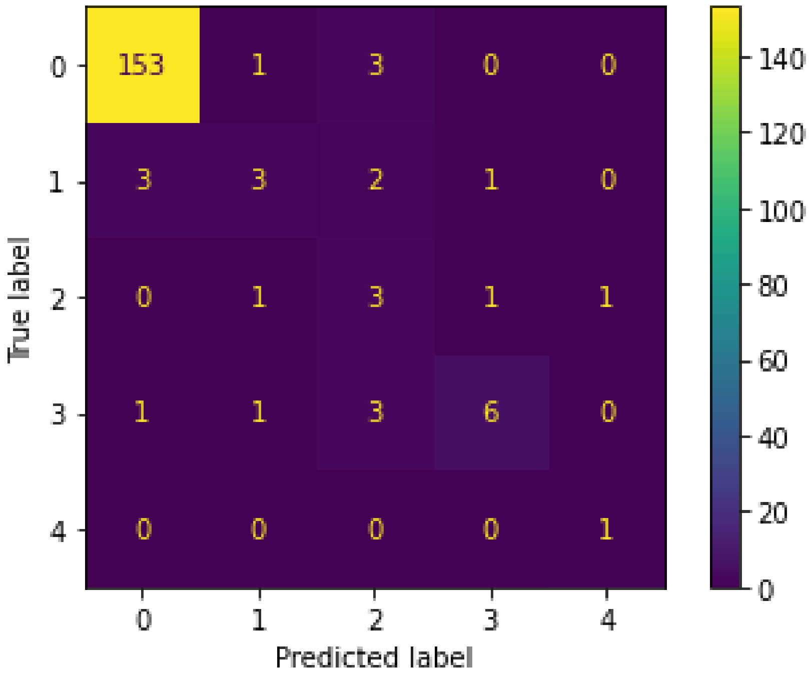


Fig. Confusion matrix for base decision tree classifier

Decision tree with stratified K-fold cross validation

K-fold cross validation is always a better way of training and evaluating a model. Its one of the variants is stratified k-fold validation in which the folds are made by preserving the percentage of samples for each class. That is each fold contains roughly the same proportions of the different class labels. It could prove to be a good technique to handle the class imbalance. So, another decision tree classifier was trained and evaluated on the 10 stratified folds of the dataset. F score of the classifier over 10 folds is plotted as below.

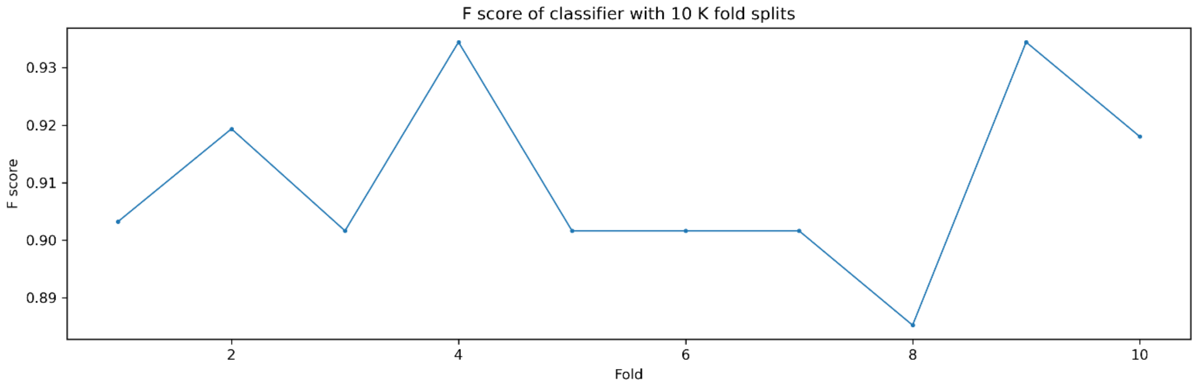


Fig. F score of decision tree classifier over 10 stratified folds

To establish that K fold stratified validation is better than normal train test splitting of data set, the experiment was done with both stratified as well as normal train test splitting. The experiment was repeated 10 times to get an average F score value. Mean F score with stratified k fold validation was 0.914, whereas mean F score with normal splitting was 0.886. Clearly, stratified K-fold cross validation is a better alternative over normal train-test splitting.

### Dealing with class imbalance: Undersampling

As already discussed, undersampling and oversampling are few of the techniques that can be used to deal with the class imbalance. An experiment was performed on undersampled data. Firstly, data is split using K-fold stratified sampling and then, training data was undersampled using clustering technique. Finally, the classifier was evaluated on the untouched test data. The experiment was repeated 10 times with different random seeds to get an average F score value. Mean F score with undersampling was 0.73, which is a very poor score comparatively. Thus, this experiment concluded that under-sampling is not a good way to deal with an imbalanced data set.

### Dealing with class imbalance: Oversampling

In a similar fashion as above, an experiment was performed on the given data set by oversampling it using SMOTE technique. Mean F score was 0.894, which was still not better as compared to the base classifier model. However, it was still better than what was achieved by undersampling.

### Random forest classifier with k fold stratified cross validation.

In this experiment, a random forest classifier was trained and evaluated using K fold stratified cross validation. First, data was split in 10 stratified folds and then, random forest classifier was trained and evaluated on it. This was repeated 10 times with different random seeds to take an average F score. Mean F score with stratified k fold was 0.924. which is the best performance achieved so far. To compare these results with the case when normal splitting would be used, a similar experiment was performed with 70-30 train test split. Mean F score with this split was 0.919. It established that random forest classifier performs better as compared to the single decision tree classifier. It also confirmed our observation so far that K fold stratified cross validation is a better way of splitting the data set into training and testing.

### Random forest classifier using oversampling with SMOTE

As we did oversampling of dataset using SMOTE technique for decision tree classifier, similar experimentation was performed for random forest classifier. Firstly, data is split using K-fold stratified sampling and then, training data was oversampling using SMOTE technique. Finally, the classifier was evaluated on the untouched test data. The experiment was repeated 10 times with different random seeds to get an average F score value. Mean F score of random forest classifier with oversampling was 0.934, which is a tremendous performance improvement over the decision tree. However, it is a bit suspicious because oversampling might have led to overfitting of the model.

Further tuning the random forest classifier

In the quest of improving the performance further, few experiments were conducted to tune the parameters of random forest algorithm like n-estimators and min-samples split. All these experiments were conducted with 10-fold stratified splitting and SMOTE oversampling in place. As we are already aware that they improve the performance significantly.

1. Experiment for n-estimators

Random forest classifier was trained with different values of n-estimators: 200, 400, 600, 80 0, 1000, 1200, 1400, 1600, 1800, 2000 and then evaluated. The performance was compared against each other. The comparison plot for the same is shown below. Random forest showed highest performance with 800 n-estimators. The highest mean F score was 0.946. This is the peak performance of the random forest classifier in all the experiments done so f ar.

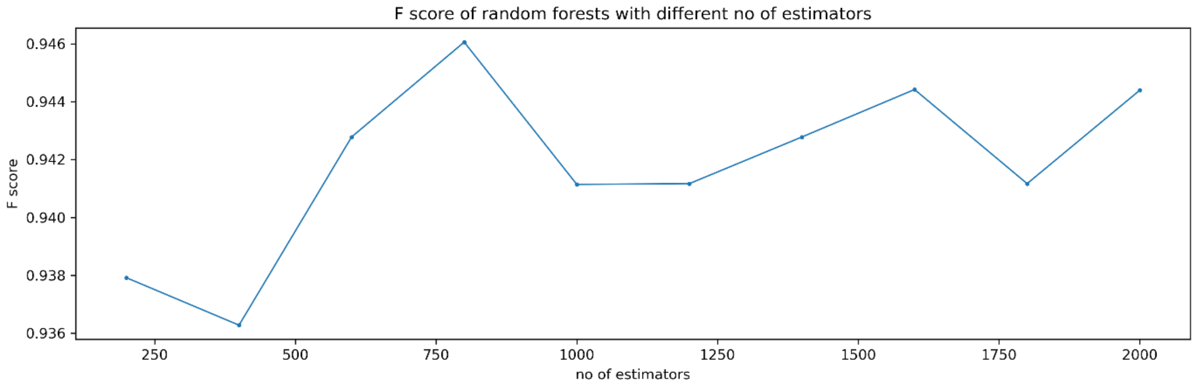


Fig. F score of random forest with different n-estimators

2. Experiment for number of minimum samples split

Random forest classifier was trained by varying values of minimum samples split. Min-sample split values used were 2,3,5,8, and 10. This was performed with 800 n-estimators as it gave highest accuracy in above experiment. Highest mean F score of 0.946 was achieved when decision trees of random forests perform splitting with 2 min samples. The figure below depicts the plot for comparison.

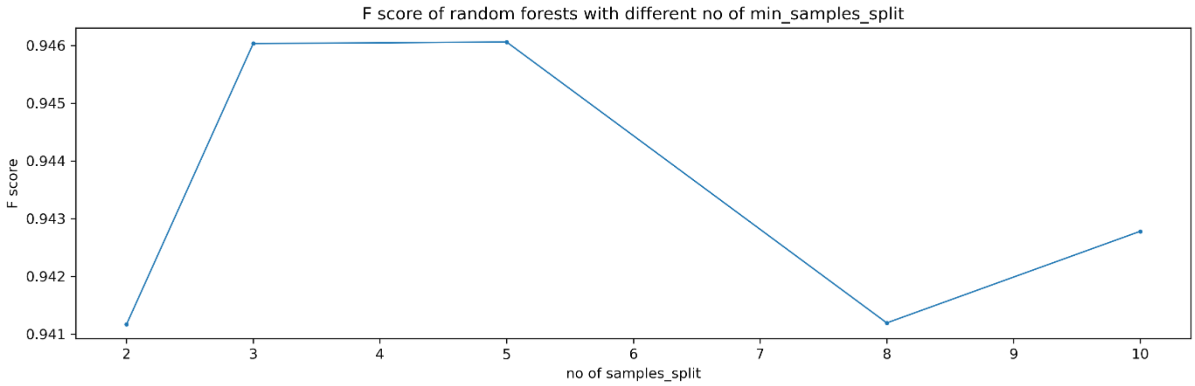


Fig.  F score comparison of random forests with different min-sample split values

## 7. Comparison of decision tree and random forest

Both decision tree and random forest classifier were trained and evaluated using combination of data splitting techniques like 70-30 train-test splitting and K-fold cross validation with sampling techniques like undersampling using clustering and oversampling using SMOTE. Using undersampling was not a good idea, however oversampling boosted the performance of both the classifiers. The experiment results indicate that random forests perform better with all the techniques. The random forest gave its best F score of 93.4%, while for decision tree the best F score was only 91.4%. The results are shown in the table for comparison. The below plot also depicts the performance comparison of the two classifiers.

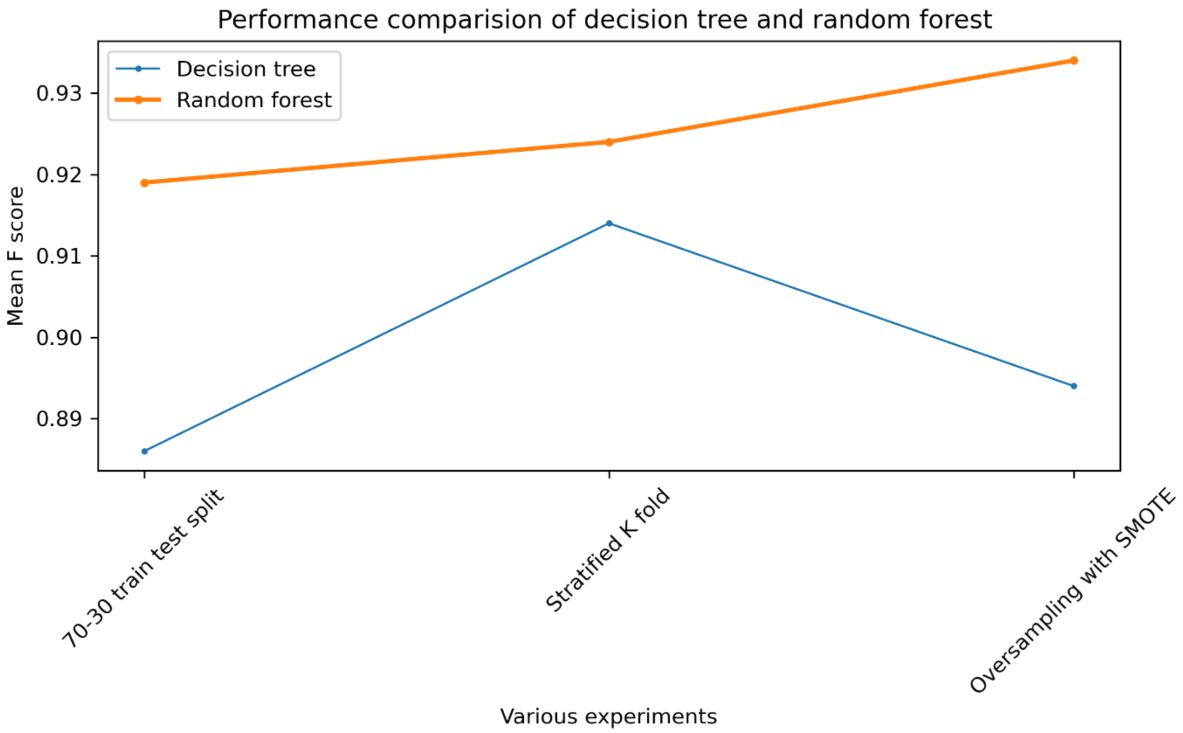


Fig. Comparison plot between random forest and decision tree

Table showing comparison (F scores) between decision tree and random forests

| F-scores | Decision tree | Random forest |
|---|---|---|
| 70-30 train test split | 0.886 | 0.919 |
| Stratified K fold | 0.914 | 0.924 |
| Under-sampling using clustering | 0.73 | NA |
| Over-sampling with SMOTE | 0.894 | 0.934 |

## 8. Conclusions and further work

This project aimed to use machine learning models, namely decision tree and random forest to predict the diagnosis of hepatitis C infection. It further goes on to classify various stages of the infection like just hepatitis, fibrosis, and cirrhosis. Although the data set has class imbalance, quite a good accuracy of 94.6 % was achieved in diagnosis of the infection by random forest. As expected, random forest performed significantly better than its decision tree counterpart. Class imbalance was dealt with up to an extent by exploiting K fold stratified cross validation and SMOTE oversampling. It would also be worth to try black box algorithms like support vector machines or deep neural networks to classify the diagnosis.

There was a constant suspicion of overfitting while training the classifiers because the data set was small, and it also had class imbalance. Further work on evaluation can be done using ROC curves, which seem fit for the case of class imbalance. The best way to improve on overfitting would be to collect more data. It would be interesting to know if dimensionality reduction has any impact on the performance of the classifiers used in the project. Removing the outliers is also a way forward. Also, further experiments can be performed on max depth limit of decision tree to keep a check on overfitting.

## References

[1] https://www.who.int/news-room/fact-sheets/detail/hepatitis-c

[2] Lichtinghagen R et al. J Hepatol 2013; 59: 236-42
Hoffmann G et al. Using machine learning techniques to generate laboratory diagnostic pathways

[3] Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. Introduction to data mining. Pearson. 2nd Edition.

[4] Gebo KA, Herlong HF, Torbenson MS, et al. Role of liver biopsy in management of chronic hepatitis C: a systematic review. Hepatology 2002;36:161-72.

## Dataset

[1] UCI: http://archive.ics.uci.edu/ml/index.php