Dissertation on

PREDICTION OF HEPATITIS AND LIVER DAMAGE

Submitted in partial fulfillment of the requirements for the award of degree of

Bachelor of Technology
in
Computer Science & Engineering

Submitted by:

M Pradeep Kumar	**01FB16ECS183**
M sumukha	**01FB16ECS185**
Sreerama Priyanka	**01FB16ECS395**

Under the guidance of

Internal Guide

Mahitha G
Professor, Department
of Computer Science
and Engineering
PES University

January – May 2020

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
FACULTY OF ENGINEERING
PES UNIVERSITY
(Established under Karnataka Act No. 16 of 2013)
100ft Ring Road, Bengaluru – 560 085, Karnataka, India

PES UNIVERSITY

(Established under Karnataka Act No. 16 of 2013)
100ft Ring Road, Bengaluru – 560 085, Karnataka, India

FACULTY OF ENGINEERING

CERTIFICATE

This is to certify that the dissertation entitled

PREDICTION OF HEPATITIS AND LIVER DAMAGE

is a bonafide work carried out by

M Pradeep Kumar	**01FB16ECS183**
M Sumukha	**01FB16ECS185**
Sreerama Priyanka	**01FB16ECS395**

In partial fulfilment for the completion of eighth semester project work in the Program of Study Bachelor of Technology in Computer Science and Engineering under rules and regulations of PES University, Bengaluru during the period Jan. 2020 – May. 2020. It is certified that all corrections / suggestions indicated for internal assessment have been incorporated in the report. The dissertation has been approved as it satisfies the 8th semester academic requirements in respect of project work.

Signature Mahitha G Professor	Signature Dr. Shylaja S S Chairperson	Signature Dr. B. K. Keshavan Dean of Faculty

External Viva

Name of the Examiners	**Signature with Date**
1. _____	_____
2. _____	_____

DECLARATION

We hereby declare that the project entitled **"*PREDICTION OF HEPATITIS AND LIVER DAMAGE*"** has been carried out by us under the guidance of *Prof. Mahitha G,* Assistant Professor, and submitted in partial fulfillment of the course requirements for the award of degree of **Bachelor of Technology** in **Computer Science and Engineering** of **PES University, Bengaluru** during the academic semester January – May 2020. The matter embodied in this report has not been submitted to any other university or institution for the award of any degree.

PLACE: Bengaluru

DATE: 30-04-2020

NAME AND SIGNATURE OF THE CANDIDATES

M Pradeep Kumar	**M Sumukha**	**S Priyanka**
(01FB16ECS183)	**(01FB16ECS185)**	**(01FB16ECS395)**

ACKNOWLEDGEMENT

We would like to acknowledge our project guide Prof. Mahitha G, for giving us constant tips and comments to progress and accomplish the project objective. Each time before review we were in constant touch with her in order to progress further. She understood our problem well and provided sources and references in order to gain the domain knowledge.

The project coordinators Prof. Preet Kanwal and Sangeeta V I always played an important role in organizing the process of demonstrating our project to the review panel by providing flexible dates. Also helped us with a suitable environment setup in order to demo the monthly project update to the panel. During the corona epidemic we were given the opportunity to demonstrate though google hangouts meetings and procedures were given in hand before through mail. The Chairperson Dr. Shylaja S played a crucial role in dividing our final credits and informing us the process of evaluation of the final semester project. We also thank the Dean of Faculty Dr. B. K. Keshavan for the whole hearted support throughout the project. We thank the Vice Chancellor Dr. Suryaprasad J for being a pillar of support throughout the project to all the students. We thank the Pro chancellor of PES University, Prof. Jawahar Doreswamy, for inspiring us with his words and speeches throughout the semester. We would also like to thank the Chairman and Chancellor, Dr. M R Doreswamy for giving us an opportunity to study in this esteemed University. We also would like to thank our family and friends who supported us constantly throughout this project. During the course of this project we have learnt a lot from various sources and are indebted to all.

ABSTRACT

Viral hepatitis is the regularly found health problem throughout the world among other easily transmitted diseases, such as tuberculosis, human immune virus, malaria and so on. Among all hepatitis viruses, the uppermost numbers of deaths are result from the long-lasting hepatitis C infection or long-lasting hepatitis B. In order to develop this system, the data is acquired through UCI machine learning repository. Once the data is acquired, it is pre-processing is done, replacing textual data with numbers and replacing missing values using imputer. Feature selection is also done using PCA (principle component Analysis) and medical research. Classification algorithms used are Decision tree, Gradient boosting, Random forest, Logistic regression and SVM. The proposed model has the ability to predict the Hepatitis and liver damage without scanning and biopsy.

TABLE OF CONTENTS

LIST OF FIGURES

6.2	Liver form	42
6.3	Hepatitis UI	43
6.4	Hepatitis form	44

1. INTRODUCTION

The Liver is the largest organ in the human body. It is found in the upper right part of the abdomen, weighing around 3 pounds. Without a healthy liver, human can't live. This is the storage and the body's filter. Whether you eat it, drink it, breathe it, or put on your skin, everything goes through the liver. Almost every cell and tissue in the body is liver-dependent. If something goes wrong with the liver, almost every other organ in the body will be severely affected. Its regenerating ability is the most impressive function of the liver. This may still develop new tissue and extend to the original proportions within weeks, as much as three-quarters of the liver can be lost. This ensures that those who need transplants will get a living donor's portion of the liver. The liver is referred to as a non-complaining organ because its cells have no nerves. Therefore, without knowing it, you may have serious liver damage. Some people feel pain in the liver region, usually caused both by the liver capsule and by surrounding organs.

Specific hepatotropic virus, which have distinct modes of distribution and epidemiology, is diffused hepatitis viral inflammatory. A viral prodrome that is unspecific is accompanied by anorexia, diarrhea, and sometimes by fever. Jaundice sometimes occurs before other symptoms normally start to resolve. Most cases are accidental, but chronic hepatitis is growing. Often acute viral hepatitis (indicating fulminant hepatitis) progresses to acute liver failure. Diagnosis includes liver function testing and serological virus testing. Good hygiene and universal treatment will avoid the acute hepatitis of the virus. Viral hepatitis is an infection that results in liver damage. This inflammation of the liver is caused by five different viruses such as hepatitis viruses A, B, C, D, and E. These five different types of viruses can cause acute disease. but the maximum number of deaths throughout the world results from chronic hepatitis virus B or hepatitis virus C infection (WHO, 2013). Chronic inflammation of the hepatitis B virus is liver inflammation, which lasts more than 6 months.

Some individuals with chronic hepatitis B are not symptomatic, but some do feel sick and exhausted. The risk of liver cancer rises with chronic hepatitis B.

Doctors diagnose hepatitis B from blood tests and also conduct a liver biopsy to see if the liver is affected. Not all chronic hepatitis B patients have to be treated, but an antiviral is administered when chronic hepatitis B affects the liver (which causes inflammation or scarring).
According to the Annual Report of WHO (2017), Hepatitis B is highly endemic and probably affects an estimated 5–8% of the population in Africa, mainly in West and Central Africa. It is estimated that around 19 million adults are infected with chronic hepatitis C in this region. As a result, nearly 2.3 million people living with HIV are also infected with hepatitis virus C and about 2.6 million people are infected with the hepatitis B virus.

This is a model where it predicts a person's liver damage and Hepatitis B and Hepatitis C. To develop this model various machine learning algorithms are used and the dataset is collected from UCI machine learning repository and feature selection is applied to it. This model takes input from lab tests and predicts the occurrence of hepatitis and whether a person's liver is damaged or not.

2. RESEARCH BACKGROUND

2.1 Literature Survey

The paper explains how the classification algorithms are applied to the acquired data set and then compares the results to each other and also predicts whether the person will live or not. Comparison and tabulation of the outcomes of the four chosen data algorithms Decision Trees, Logistic Regression, Linear Support Vector Machine. The results have concluded as follows: with the optimum accuracy of 87.17% and with the optimum accuracy of the Decision Tree algorithm, 82.05%. The logistic regression algorithm provides the following detail. The Linear Support Vector Machine with optimal accuracy of 76.92% is based on a decision-making tree algorithm and finally the Naïve Bayes algorithm with optimal accuracy of 76.92% [1].

.

Further Enhancement: More research will further increase the accuracy of all these classification algorithms by increasing the data set size and by using ensemble methods such as bagging and boosting. Besides, the process of classification of the dataset can be automated by creating a simple UI.

Many studies have shown the ability of Machine Learning and Data Mining tools in the field of medicine to identify secret predictive trends from the medical databases, providing high diagnostic potential for disease. Therefore, aim is early detection of the stage of fibrosis in Egyptian Hepatic C patients and to select the major parameters from individual laboratory tests. As a computer teaching strategy, Decision tree method have been applied.

Each biomarker detects the level of fibrosis. These serum markers can replace hepatic biopsy. The interpretation of the phases of hepatic fibrosis is primarily classified into five different phases• F0 = no fibrosis • F1 = portal fibrosis septal • F2 = few fibrosis septal • F3= numerous septal fibrosis • F4 = cirrhosis

Classification is the issue of which category/set of categories a new case is based on the problem field and included data. Classification in machine learning is like finding new sets by learning from past

experiences. Decision tree classification method have been applied in the categorization problem for classifying and predicting fibrosis grades (F0, F1, F2, F3, F4) [2].

Further enhancement:

The problem of classification is to classify which category/set of categories a new case is based on the problem domain and the data included. Classification in machine learning involves finding new sets through learning from old examples. Instantaneously Decision Treat classification technique have been used in the classification question for the classification and estimation of fibrosis degrees (F0, F1, F2, F3, F4).

Identify whether the patient has a liver disorder or not is the primary purpose of this research. Many of these criteria are used to assess the state of the liver and to compare the effects of various techniques for decision tree. Weka has been developed at Waikato and is an information mining tool in java. WEKA is a highly effective data mining platform for accuracy classifying with the application of various algorithmic methods, as well as comparing it with datasets. The comparison of different algorithms from the decision tree is made. It is seen that Decision Stump is the best classifier in this case, and the accuracy of the algorithm is 70.67%. This figure indicates that the liver disease of a new patient is accurately estimated to have a 70.67 percent ratio. The accuracy rate of the random tree, LMT is and hefting tree is 69.47 percent, 69.0%, and 69.75 percent [3].

The attributes and structure of the research database – attributes and the database structure were closely analyzed to identify useful attributes from the patient database of the liver.

Determine the essence and description of the question of study and workflow to obtain exact and optimal results. An ILPD (Indian Liver Patient Dataset)-UCI Dataset analysis of datasets.

Arrange and assign useful attributes to the database and evaluate them using the WEKA method to generate the results.

Further enhancement:

This paper outlines a method used for developing the community health services hybrid model.

Among other dominant illnesses, like cardiac and diabetes prediction and diagnosis, these diagnosis algorithms may also be applied. The goal is to see how the environment is changed with the implementation of modern algorithms to potential techniques.[4]

3. METHODOLOGY

3.1. Data Pre-processing

Data pre-processing, it is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviours or trends, and is likely to contain many errors. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis.

3.1.1 Hepatitis B and Liver

As ML models won't accept textual data so yes/no values have been replaced with numbers. Missing values have also been replaced using imputer.

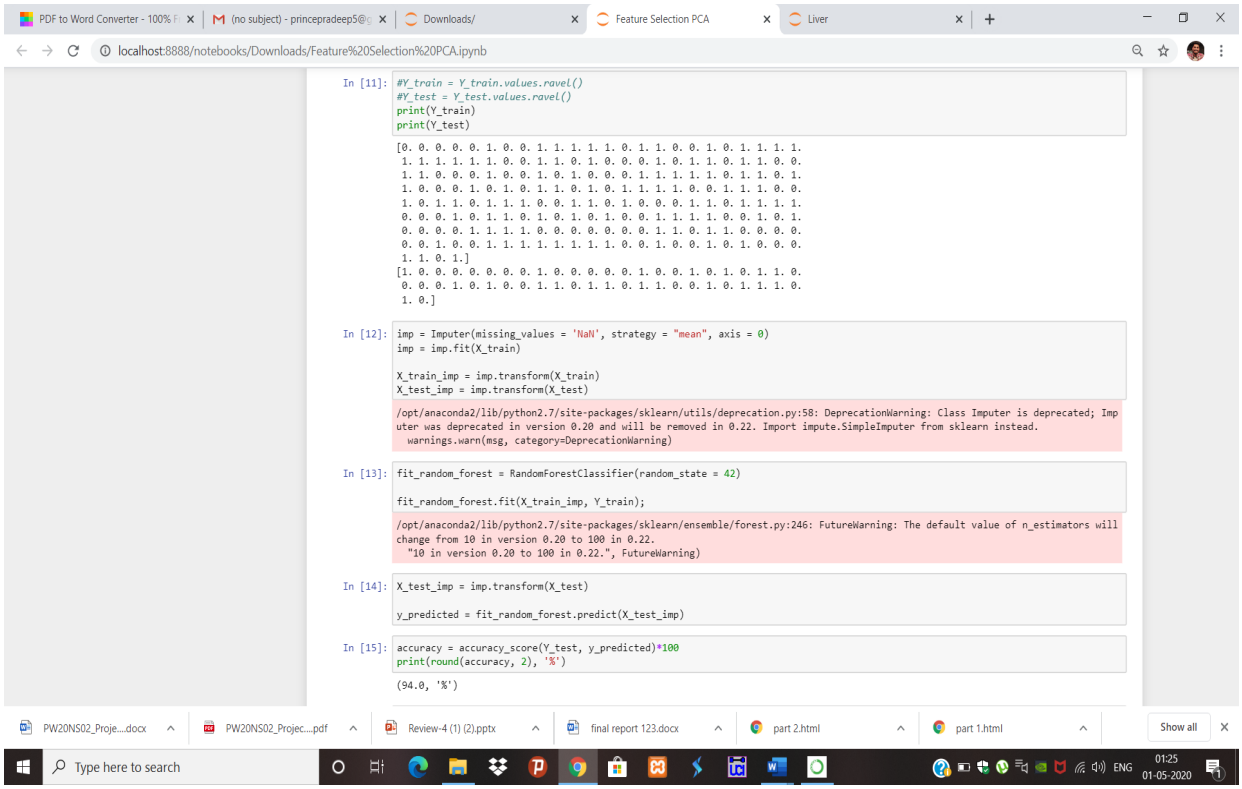


Figure3.1 : Imputer code for missing values

In the **figure** imputer is used to replace missing values. Once the data set is divided into training set(X_train) and (X_test) imputer is applied. The missing_data placeholder which has to be imputed. By default is NaN. The data which will replace the NaN values is from the dataset. The strategy argument can take the values – 'mean'(default). After replacing X_train is converted to X_train_imp and X_test is converted to X_test_imp.

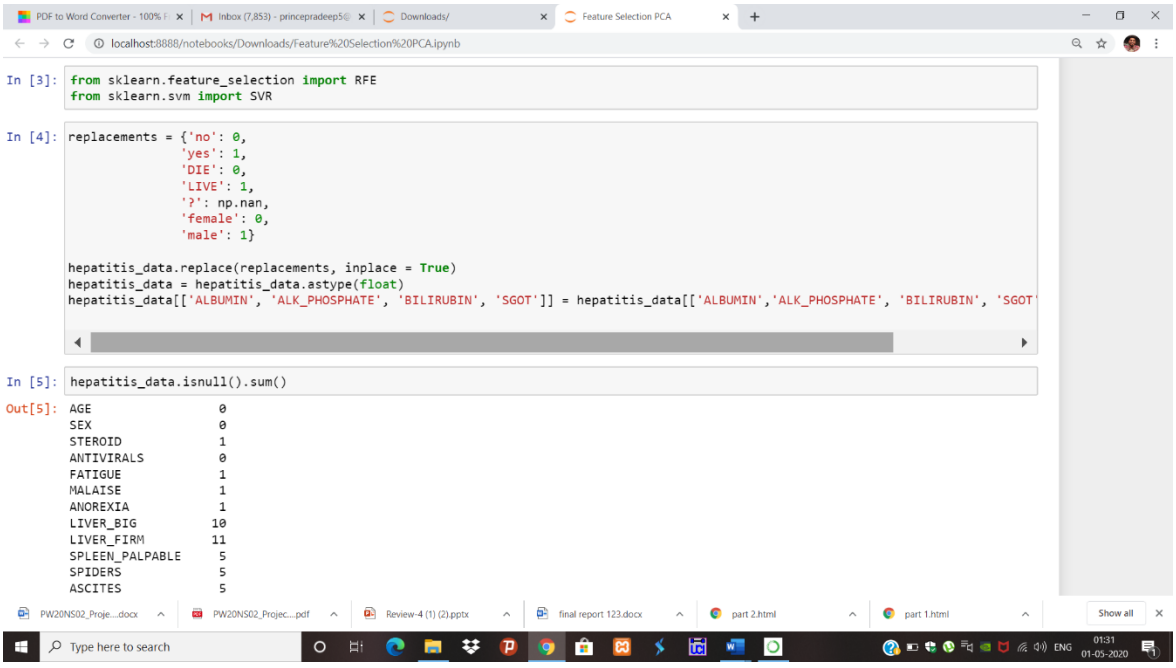


Figure 3.2: Replacement of textual data

In figure 2 textual data has been replaced with numbers as textual data is not acceptable in ML. In the dataset there are yes or no values, live die values and female male values. So yes is replaced with "1"and no with "0". Live is replaced with "1" and die with "0". Male is replaced with "1" and female with "0".

3.1.2 Hepatitis C :

In hepatitis C dataset, it did not contain any missing value or textual values. So above methods were not implemented. By Looking at the Hepatitis C dataset, It can be observed that there is a huge difference between attribute values. Few attributes range is different compared to other attributes. If we consider ALT and RNA Base, ALT value ranges in hundreds but for RNA Base it ranges in Lakhs. It will be hard for Machine Learning model to learn this kind of dataset with Limited number of rows.

	Age	Gender	BMI	Fever	Nausea/Vomting	Headache	Diarrhea	Fatigue & generalized bone ache	Jaundice	Epigastric pain	...	ALT 36	ALT 48	ALT after 24 w	RNA Base	RNA 4	RNA 12	RNA EOT
0	56	1	35	2	1	1	1	2	2	2	...	5	5	5	655330	634536	288194	5
1	46	1	29	1	2	2	1	2	2	1	...	57	123	44	40620	538635	637056	336804
2	57	1	33	2	2	2	2	1	1	1	...	5	5	5	571148	661346	5	735945
3	49	2	33	1	2	1	2	1	2	1	...	48	77	33	1041941	449939	585688	744463
4	59	1	32	1	1	2	1	2	2	2	...	94	90	30	660410	738756	3731527	338946

mting	Headache	Diarrhea	Fatigue & generalized bone ache	Jaundice	Epigastric pain	...	ALT 36	ALT 48	ALT after 24 w	RNA Base	RNA 4	RNA 12	RNA EOT	RNA EF	Baseline histological Grading	Baselinehistological staging
1	1	1	2	2	2	...	5	5	5	655330	634536	288194	5	5	13	2
2	2	1	2	2	1	...	57	123	44	40620	538635	637056	336804	31085	4	2
2	2	2	1	1	1	...	5	5	5	571148	661346	5	735945	558829	4	4
2	1	2	1	2	1	...	48	77	33	1041941	449939	585688	744463	582301	10	3
1	2	1	2	2	2	...	94	90	30	660410	738756	3731527	338946	242861	11	1

Figure 3.3: Hepatitis C Dataset

Now each attribute value is grouped based on clinical laboratory range for that attribute. It can be seen as shown below,

```
In [8]: def Age(val):
            if(val<32):
                return 1
            elif(val<37):
                return 2
            elif(val<42):
                return 3
            elif(val<47):
                return 4
            elif(val<52):
                return 5
            elif(val<57):
                return 6
            elif(val<62):
                return 7
            else:
                return 8
```

Figure 3.4: Pre-processing Age attribute

```
In [10]: def BMI(val):
             if(val<18.5):
                 return 1
             elif(val<25):
                 return 2
             elif(val<30):
                 return 3
             elif(val<35):
                 return 4
             else:
                 return 5
```

Figure 3.5: Pre-processing Body Mass Ratio attribute

```
def WBC(val):
    if(val<4000):
        return 1
    elif(val<11000):
        return 2
    elif(val<12101):
        return 3
    else:
        return 4
```

Figure 3.6: Pre-processing WBC attribute

```
def RBC(val):
    if(val<3000000):
        return 1
    elif(val<5000000):
        return 2
    elif(val<5018451):
        return 3
    else:
        return 4
```

Figure 3.7: Pre-processing RBC attribute

```
def HGB(sex,val):
    if(sex == 1):
        if(val<14):
            return 1
        elif(val<17.5):
            return 2
        elif(val<20):
            return 3
        else:
            return 4
    else:
        if(val<12.3):
            return 1
        elif(val<15.3):
            return 2
        elif(val<20):
            return 3
        else:
            return 4
```

Figure 3.8: Pre-processing HGB attribute

```
def Plat(val):
    if(val<100000):
        return 1
    elif(val<255000):
        return 2
    elif(val<226465):
        return 3
    else:
        return 4
```

Figure 3.9: Pre-processing Platelet count attribute

```
def RNA(val):
    if(val<=5):
        return 1
    else:
        return 2
```

Figure 3.10: Pre-processing RNA attributes

```
In [9]: def ASL_AST(val):
            if(val<20):
                return 1
            elif(val<40):
                return 2
            elif(val<128):
                return 3
            else:
                return 4
```

Figure 3.11: Pre-processing AST and ALT attributes

In [33]:

```
for i in range(len(df)):
    df.iloc[i,0]=Age(df.iloc[i,0])
    df.iloc[i,2]=BMI(df.iloc[i,2])
    df.iloc[i,10]=WBC(df.iloc[i,10])
    df.iloc[i,11]=RBC(df.iloc[i,11])
    df.iloc[i,12]=HGB(df.iloc[i,1],df.iloc[i,12])
    df.iloc[i,13]=Plat(df.iloc[i,13])
    df.iloc[i,14]=ASL_AST(df.iloc[i,14])
    df.iloc[i,15]=ASL_AST(df.iloc[i,15])
    df.iloc[i,16]=ASL_AST(df.iloc[i,16])
    df.iloc[i,17]=ASL_AST(df.iloc[i,17])
    df.iloc[i,18]=ASL_AST(df.iloc[i,18])
    df.iloc[i,19]=ASL_AST(df.iloc[i,19])
    df.iloc[i,20]=ASL_AST(df.iloc[i,20])
    df.iloc[i,21]=ASL_AST(df.iloc[i,21])
    df.iloc[i,22]=RNA(df.iloc[i,22])
    df.iloc[i,23]=RNA(df.iloc[i,23])
    df.iloc[i,24]=RNA(df.iloc[i,24])
    df.iloc[i,25]=RNA(df.iloc[i,25])
    df.iloc[i,26]=RNA(df.iloc[i,26])
```

Figure 3.12: Replacing Every attribute with pre-processed value

After pre-processing attributes contain values of similar range. Which might enable Machine Learning model to train better. Pre-processed data can be seen below.

	Age	Gender	BMI	Fever	Nausea/Vomting	Headache	Diarrhea	\
0	6	1	5	2	1	1	1	
1	4	1	3	1	2	2	1	
2	7	1	4	2	2	2	2	
3	5	2	4	1	2	1	2	
4	7	1	4	1	1	2	1	
5	7	2	2	2	2	2	1	
6	4	2	3	1	1	2	2	
7	5	2	4	1	1	2	2	
8	4	1	2	1	1	2	2	
9	4	1	4	2	1	2	2	
10	3	2	2	2	1	2	1	

	Fatigue & generalized bone ache	Jaundice	Epigastric pain	WBC	RBC	\
0	2	2	2	2	2.0	
1	2	2	1	4	2.0	
2	1	1	1	2	2.0	
3	1	2	1	2	2.0	
4	2	2	2	1	2.0	
5	2	2	1	3	2.0	
6	2	2	2	3	2.0	
7	1	1	2	2	2.0	
8	2	1	2	2	2.0	
9	1	1	2	2	2.0	
10	2	2	1	2	2.0	

	HGB	Plat	RNA Base	RNA 4	RNA 12	RNA EOT	RNA EF	\
0	2	2.0	2	2	2	1	1	
1	1	2.0	2	2	2	2	2	
2	1	2.0	2	2	1	2	2	
3	1	2.0	2	2	2	2	2	
4	1	2.0	2	2	2	2	2	
5	2	2.0	2	2	1	1	1	
6	1	2.0	2	2	2	2	2	
7	1	2.0	2	2	2	2	2	
8	1	2.0	2	2	1	2	2	
9	1	1.0	2	2	2	2	2	
10	1	2.0	2	2	2	2	2	

	Baseline histological Grading
0	13
1	4
2	4
3	10
4	11
5	4
6	12
7	12
8	5
9	4
10	15

Figure 3.13: Hepatitis C dataset after pre-processing

3.2 Prediction Models used:

a. Random Forest

Random forest is an ensemble learning method for regression, classification, and other tasks. It works by constructing a multitude of decision trees at training time. Output is the mean prediction of individual trees.

Random forests can accommodate missing values. Hepatitis B and Liver damage dataset contains missing values so this algorithm has been implemented in those datasets. By applying pre-processing and feature selection random forest algorithm is applied on the Hepatitis C dataset. **Random Forest** increases predictive power of the algorithm and also helps prevent overfitting. As shown in

figure 1 once imputer is applied for missing values random forest classifier is applied to prevent overfitting.

b. Decision Trees

Decision Tree uses a tree-like model of decisions and possible consequences, including event outcomes, resource costs, and utility. This algorithm contains only conditional control statements.

The Presence of an enzyme or symptom can signify disease or damage presence. So this algorithm is implemented in Hepatitis B, Hepatitis C, and Liver Damage predictions. As there is a chance of overfitting in Decision tree the Oversampling technique SOMTE is not applied for Hepatitis B prediction. Since Hepatitis B contains less number of rows in the dataset, oversampling led to overfitting.

c. Gradient Boosting:

This a Machine Learning Technique which suits for classification and regression problems. The prediction model is produced in the form of an ensemble of weak prediction models. Typically Decision Tree is used as weak prediction models.

Gradient Boosting Technique is used for Hepatitis B disease prediction after pre-processing. This model is an ensemble of weak prediction models. Gradient boosting involves creating and adding trees to model. Learning rates used for increasing trees are [0.05, 0.1, 0.25, 0.5, 0.75, 1]. Higher learning rates results in overfitting so “0.5” is optimal for this model. n_estimators represent the number of trees in the forest. Usually higher the number of trees the better to learn the data. However, adding a lot of trees can slow down the training process considerably. Increasing the number of estimators may result in overfitting also. In this model, using 20 trees is optimal. max_features represents the number of features to consider when looking for the best split. Increasing max_features to consider all of the features results in an overfitting in this case. Using max_features = 2 seems to get the optimal performance.

d. SVC:

This a non-parametric clustering algorithm that doesn't make any assumptions on the shape or number of the clusters in the dataset. SVC works best for low dimensional dataset. This requires a pre-processing or PCA or feature selection if the dimension of the data is high.

SVC is applied for Liver Damage Prediction without using feature elimination as several attributes are 10. In the case of Hepatitis C as there are 30 attributes, features were eliminated by PCA, RFE, and Hepatitis disease domain knowledge. Then the SVC technique is applied for the Hepatitis C dataset.

e. Logistic Regression:

Logistic Regression is a supervised **learning** classification algorithm used to predict the probability of a target variable. Logistic Regression analysis is appropriate to use when the dependent variable is dichotomous. Only the meaningful variables should be included. The independent variables should be independent of each other. For this reason Logistic Regression technique is used for Hepatitis B and Liver damage prediction. In hepatitis. In hepatitis C as there are few attributes, ALT(Alanine transferase) and RNA(ribonucleic acid) which signifies the presence or absence of symptom. Due to this Logistic regression is applied for Hepatitis C too.

3.3 Attributes

3.3.1 Attributes for dataset-Liver

AGE	Age of the patient
GENDER	Gender of the patient
TOTAL BILIRUBIN	Yellow breakdown product of normal RBC in liver and secreted in bile.
DIRECT BILIRUBIN	Water soluble and is made by the liver from bilirubin
ALKALINE_PHOSPHOTASE	Liver enzyme
ALAMINE_AMINOTRANSFERASE	Liver enzyme
ASPARTATE_AMINOTRANSFERASE	Liver enzyme
TOTAL_PROTEIN	Total amount of albumin and globulin in the body.
ALBUMIN	Total amount of albumin in the body.
ALBUMIN_AND_GLOBULIN_RATIO	1 (LESS THAN 2)

```
('Number of patients diagnosed with liver Damage: ', 416)
('Number of patients not diagnosed with liver Damage: ', 167)
```

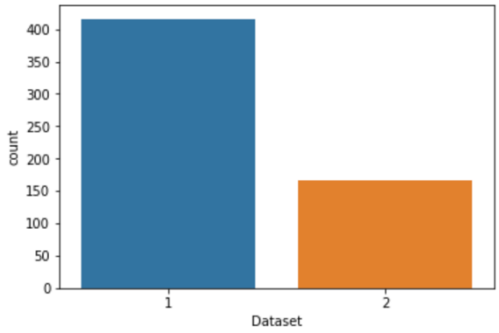


Figure 3.14: Histogram displaying Patients without Liver Damage and with Liver Damage

```
('Number of patients that are male: ', 441)
('Number of patients that are female: ', 142)
```

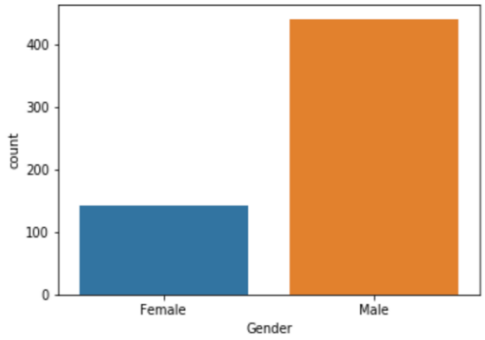


Figure 3.15: Histogram displaying male and Female count in the dataset

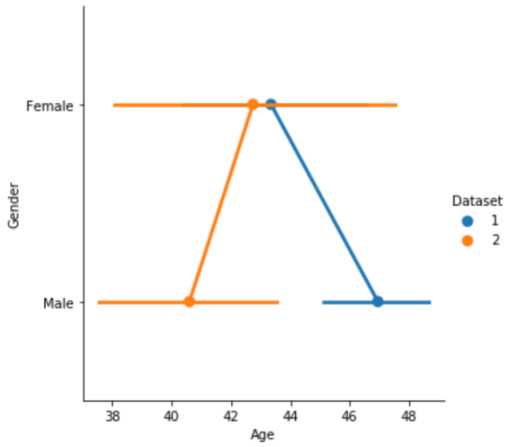


Figure 3.16: Representing relationship between Age and gender

:

	Dataset	Gender	Age
2	2	Female	50
3	2	Male	117
0	1	Female	92
1	1	Male	324

Figure 3.17: The Age count of each gender according to target class

	Dataset	Gender	Age
2	2	Female	42.740000
3	2	Male	40.598291
0	1	Female	43.347826
1	1	Male	46.950617

Figure 3.18: The mean of Age calculated for each gender.

By using figure 3.18 result, the mean for each gender according to the targeted column is calculated.

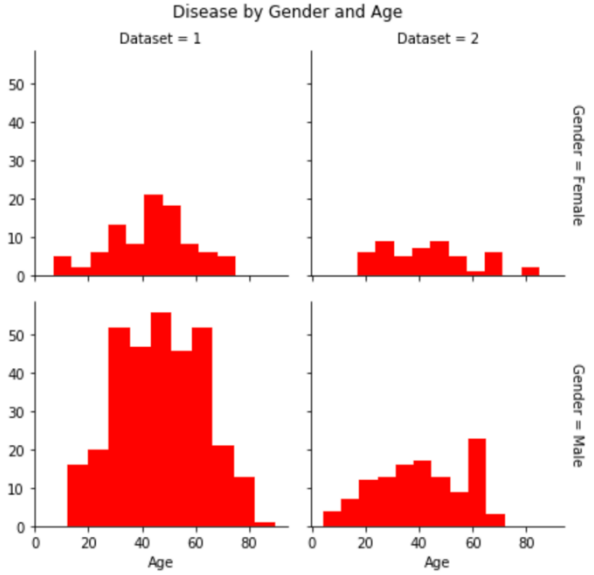


Figure 3.19: Liver Damage according to the gender

By observing the above figure, It can be noticed that male patients are highly affected by liver damage compared to female patients.

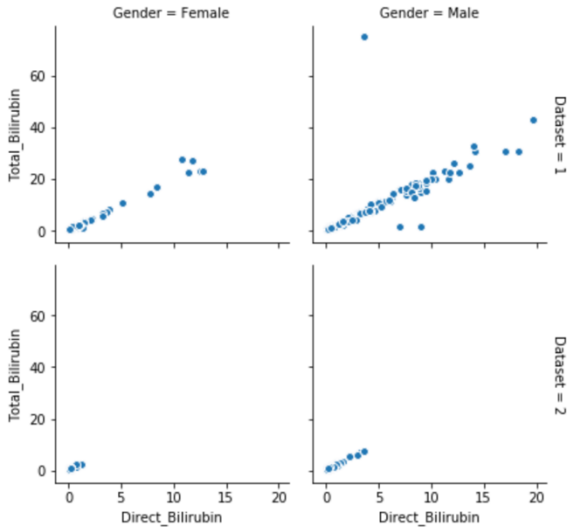


Figure 3.20: Relationship between Total Bilirubin and Direct Bilirubin

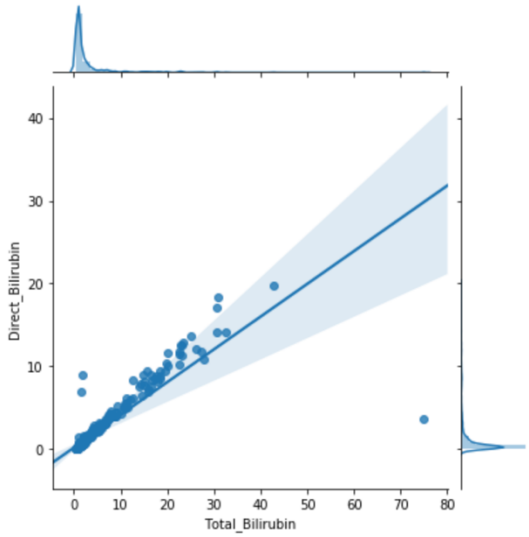


Figure 3.21: Relationship between Total Bilirubin and Direct Bilirubin.

From figure 3.20 and 3.21, It shows existence of direct relationship between Direct Bilirubin and Total Bilirubin. (Note: Direct Bilirubin is formed when Bilirubin undergoes some chemical change in the liver).

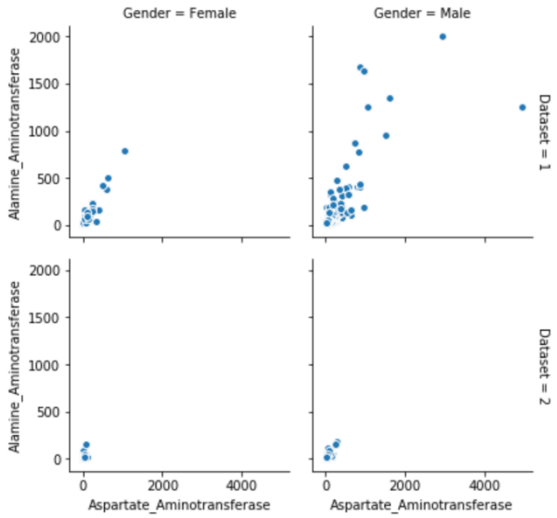


Figure 3.22: Relationship between Aspartate Aminotransferase and Alanine Aminotransferase

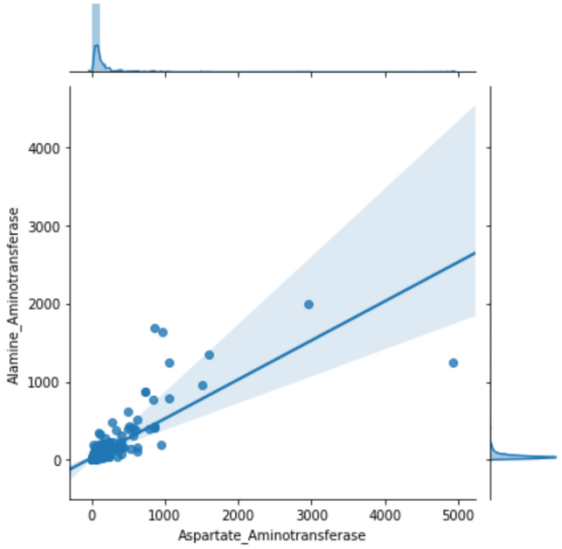


Figure 3.23:Relationship between Aspartate Aminotransferase and Alanine Aminotransferase.

From above plots It can be noticed that there is a direct Relationship between Aspartate Aminotransferase and Aspartate Aminotransferase.

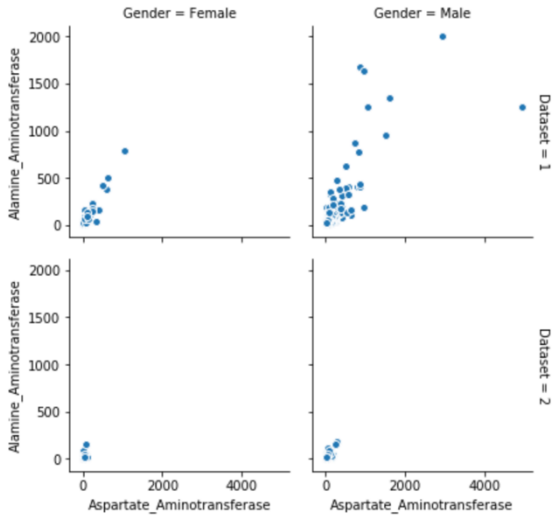


Figure 3.24: Relationship between Alkaline Phosphatase and Aminotransferase.

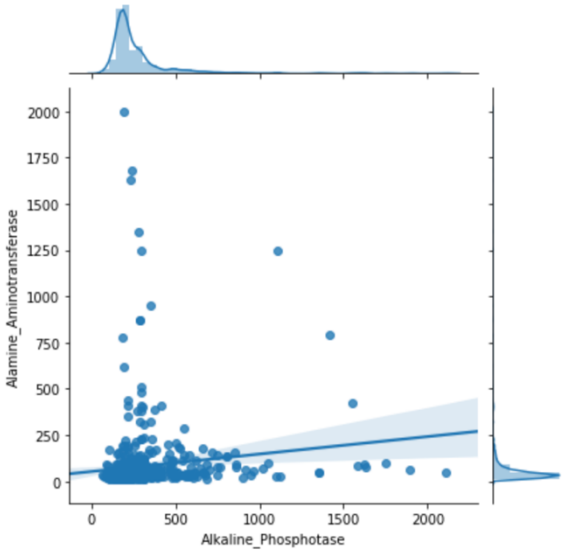


Figure 3.25: Relationship between Alkaline Phosphatase and Aminotransferase.

It can be noticed that there is no linear Relationship between Alkaline Phosphatase and Aminotransferase.

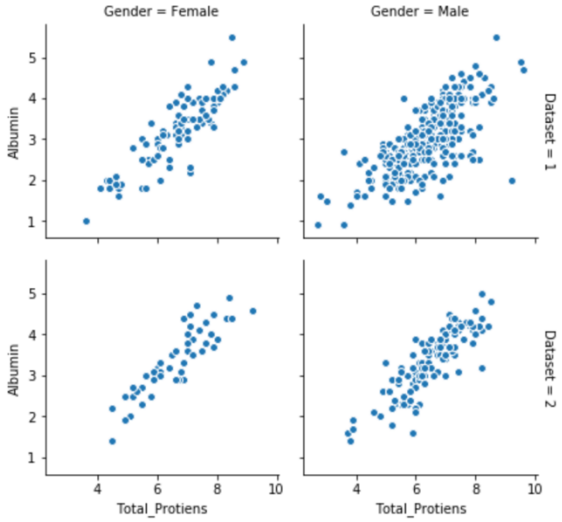


Figure 3.26: Relationship between Total Proteins and Albumin and gender.

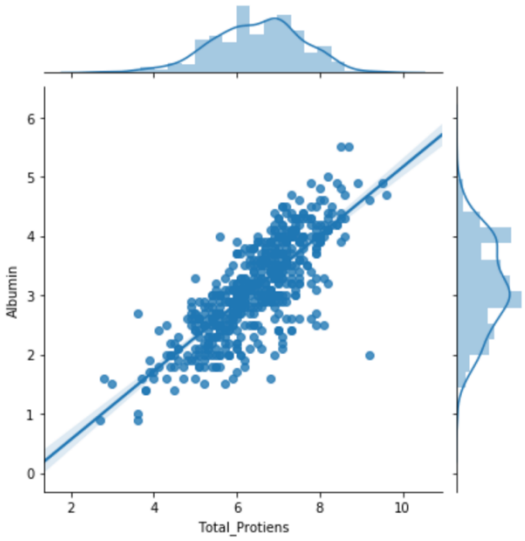


Figure 3.27: Relationship between Total Proteins and Albumin.

From figure 3.26 and 3.27, It helps to notice linear relationship between Total Proteins, Albumin and Gender.

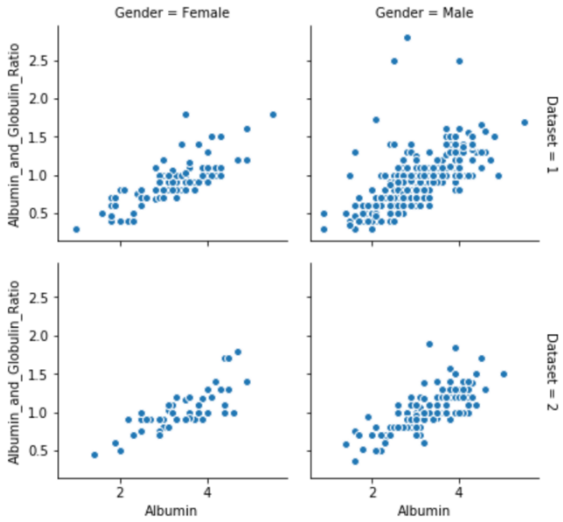


Figure 3.28: Relationship between Albumin and Globulin Ration, Albumin, and Gender.

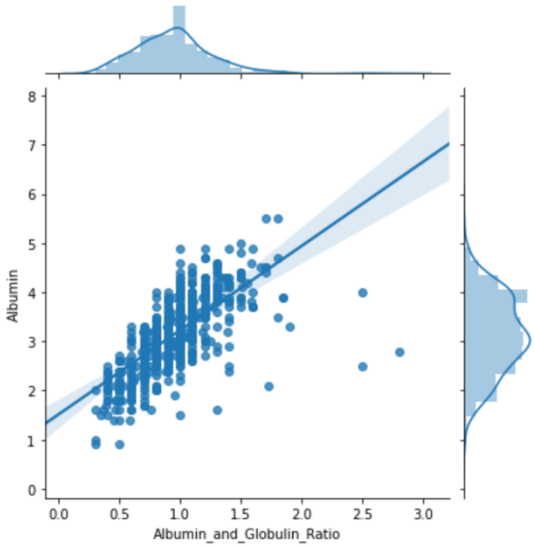


Figure 3.29: Relationship between Albumin and Globulin Ration and Albumin.

From figure 3.28 and 3.29, the linear relationship between Albumin and Globulin Ratio and Albumin can be observed.

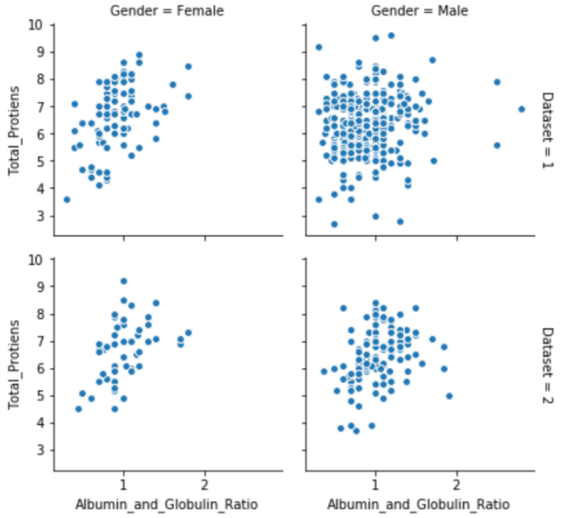


Figure 3.30: Relationship between Albumin and Globulin Ration and Total Proteins.

It can be noticed from the above figure that there is no linear relationship between Albumin and Globulin Ration, and Albumin.

3.3.2 Attributes for dataset-Hepatitis B

CLASS	If the person lives or die
AGE	Age of the patient
SEX	Gender of patient
STEROIDS	Consumption of steroids
ANTIVIRALS	Drugs use in hepatitis
FATIGUE	Tiredness
MALAISE	Discomfort
PROTIME	Blood test used to measure the amount of time required to clot.

HISTOLOGY	Study of anatomy of micro tissues through microscope
ANOREXIA	Loss of appetite
LIVER BIG	Enlargement of liver
LIVER FIRM	If the liver is firm or not
SPLEEN PALPABLE	Enlargement of spleen at least twice of its size to be palpable
SPIDERS	Type of swollen blood vessels which radiate like spider web beneath skin seen in liver disease
ASCITES	Accumulation of fluid in the abdominal cavity most often related to ascites
VARICES	Abnormal veins in the esophagus due to obstruction of blood flow to liver
BILIRUBIN	Yellowish substance in blood stream
ALK PHOSPHATE	Liver enzyme
SGOT	Liver enzyme
ALBUMIN	Liver Protein

	AGE	SEX	STEROID	ANTIVIRALS	FATIGUE	MALAISE	ANOREXIA	LIVER_BIG	LIVER_FIRM	SPLEEN_PALPABLE	SPIDERS	ASCITES	VARICES	BILIRUB
0	30	male	no	no	no	no	no	no	no	no	no	no	no	
1	50	female	no	no	yes	no	no	no	no	no	no	no	no	0
2	78	female	yes	no	yes	no	no	yes	no	no	no	no	no	0
3	31	female	?	yes	no	no	no	yes	no	no	no	no	no	0
4	34	female	yes	no	no	no	no	yes	no	no	no	no	no	

IA	LIVER_BIG	LIVER_FIRM	SPLEEN_PALPABLE	SPIDERS	ASCITES	VARICES	BILIRUBIN	ALK_PHOSPHATE	SGOT	ALBUMIN	PROTIME	HISTOLOGY	Class
no	no	no	no	no	no	no	1	85	18	4	?	no	LIVE
no	no	no	no	no	no	no	0.9	135	42	3.5	?	no	LIVE
no	yes	no	no	no	no	no	0.7	96	32	4	?	no	LIVE
no	yes	no	no	no	no	no	0.7	46	52	4	80	no	LIVE
no	yes	no	no	no	no	no	1	?	200	4	?	no	LIVE

Figure 3.31: Hepatitis B Dataset values

This is how the dataset for hepatitis B which contains text yes/no representing presence or absence of symptom or enzyme.

```
In [9]: total_of_patients = hepatitis_data.shape[0]
        total_of_live_patients = (np.sum(hepatitis_data['Class'] == 1)/total_of_patients)*100
        total_of_dead_patients = (np.sum(hepatitis_data['Class'] == 0)/total_of_patients)*100
        print("Living patients:", round(total_of_live_patients,2),"%")
        print("Dead patients:", round(total_of_dead_patients,2),"%")

        Living patients: 79.35 %
        Dead patients: 20.65 %
```

Figure 3.32: Represents percentage of Living and Dead patients in the dataset

In Hepatitis B dataset 79.35% of people are alive.

```
In [10]: numerical_variables = ['AGE', 'BILIRUBIN', 'PROTIME', 'ALBUMIN', 'ALK_PHOSPHATE', 'SGOT']
         hepatitis_data[numerical_variables].describe()
```

Out[10]:

	AGE	BILIRUBIN	PROTIME	ALBUMIN	ALK_PHOSPHATE	SGOT
count	155.000000	149.000000	88.000000	139.000000	126.000000	151.00000
mean	41.200000	1.427517	61.852273	3.817266	105.325397	85.89404
std	12.565878	1.212149	22.875244	0.651523	51.508109	89.65089
min	7.000000	0.300000	0.000000	2.100000	26.000000	14.00000
25%	32.000000	0.700000	46.000000	3.400000	74.250000	31.50000
50%	39.000000	1.000000	61.000000	4.000000	85.000000	58.00000
75%	50.000000	1.500000	76.250000	4.200000	132.250000	100.50000
max	78.000000	8.000000	100.000000	6.400000	295.000000	648.00000

Figure 3.33: summarizing the central tendency and dispersion

Here the descriptive statistics for Hepatitis B dataset is generated by excluding missing values. All attributes are not used as most of the textual data.

```
In [11]: categorical_variables = ['SEX' , 'STEROID', 'ANTIVIRALS', 'FATIGUE','MALAISE','ANOREXIA','LIVER_BIG','LIVER_FIRM','SPLE
         'SPIDERS','ASCITES','VARICES', 'HISTOLOGY']
         hepatitis_data[categorical_variables].apply(pd.Series.value_counts)
```

Out[11]:

	SEX	STEROID	ANTIVIRALS	FATIGUE	MALAISE	ANOREXIA	LIVER_BIG	LIVER_FIRM	SPLEEN_PALPABLE	SPIDERS	ASCITES	VARICES	HISTOLOGY
0.0	139	76	131	54	93	122	25	84	120	99	130	132	85
1.0	16	78	24	100	61	32	120	60	30	51	20	18	70

Figure 3.34: count of each class (0 or 1 corresponds to 'no' and 'yes').

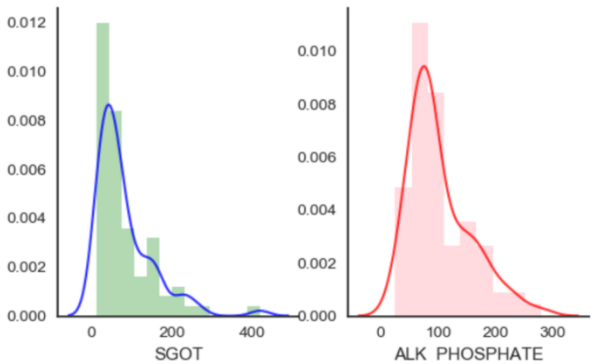


Figure 3.35: Histogram for SGOT and ALK PHOSPATE

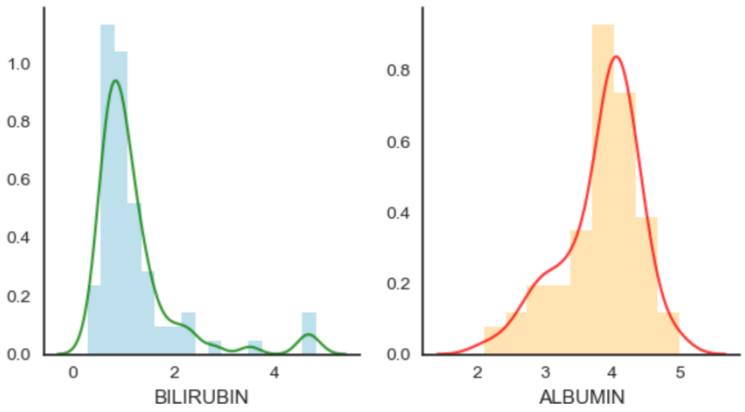


Figure 3.36: Histogram for BILIRUBIN and ALBUMIN

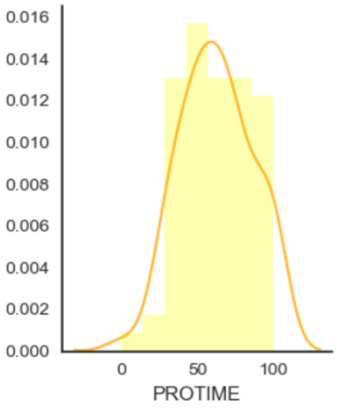


Figure 3.37: Histogram for PROTIME

This shows that several variables show degree of skewness. To fix this we applied log-transform on variables.

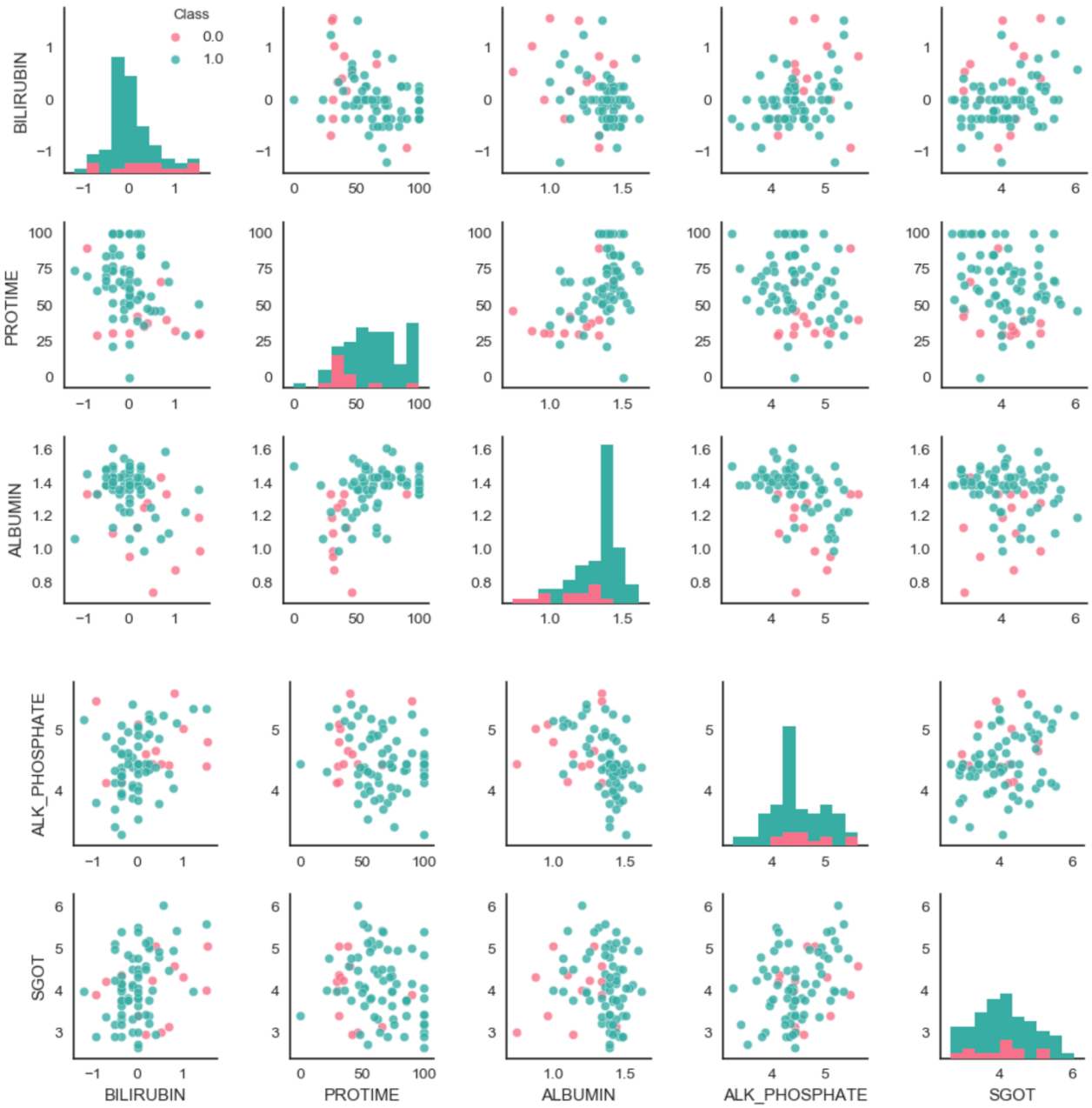
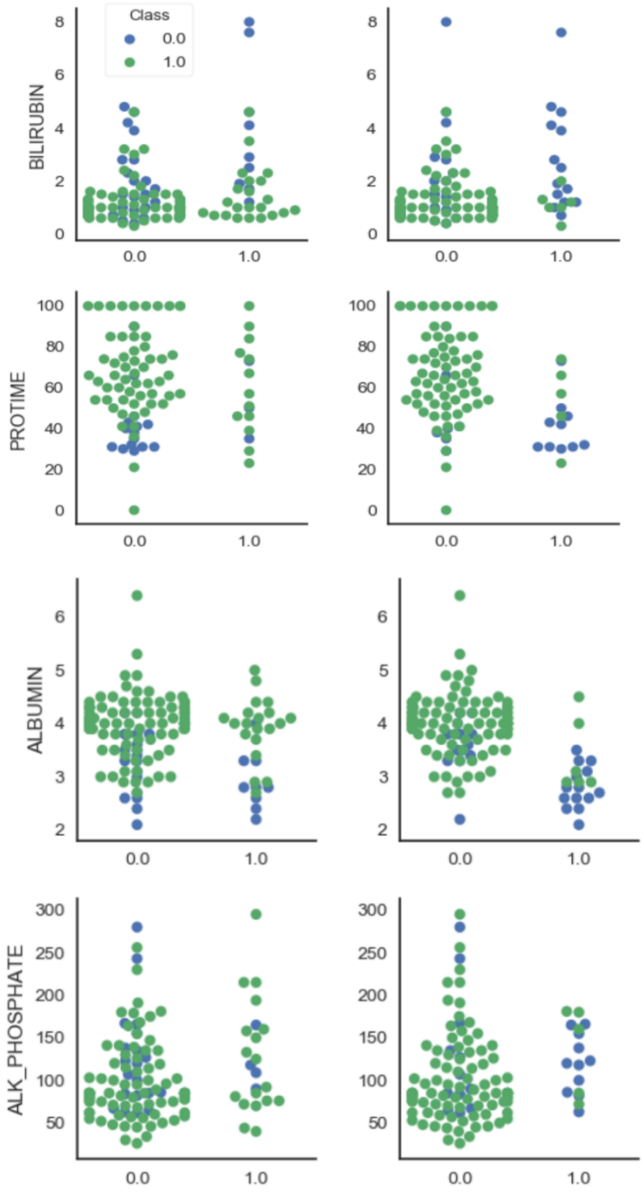


Figure 3.38: Relationship between different numerical variables

Pair plot function is used to visualize the relationship between different numerical values. Patients in class 0 and 1 are being observed which define survival of patient. There is also no linear relationship between the variables plotted. Few variables can be differentiated to which class they belong but the distinction is not clear.



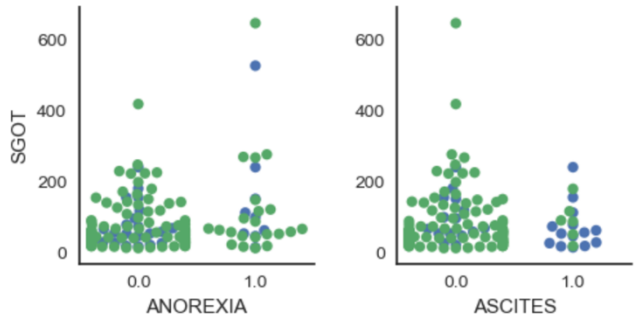


Figure 3.39: Relation between categorical values and numerical values.

Variables plotted for anorexia showed no difference and even for ascites. It can be observed that patients with class 0 tend to have ascites.

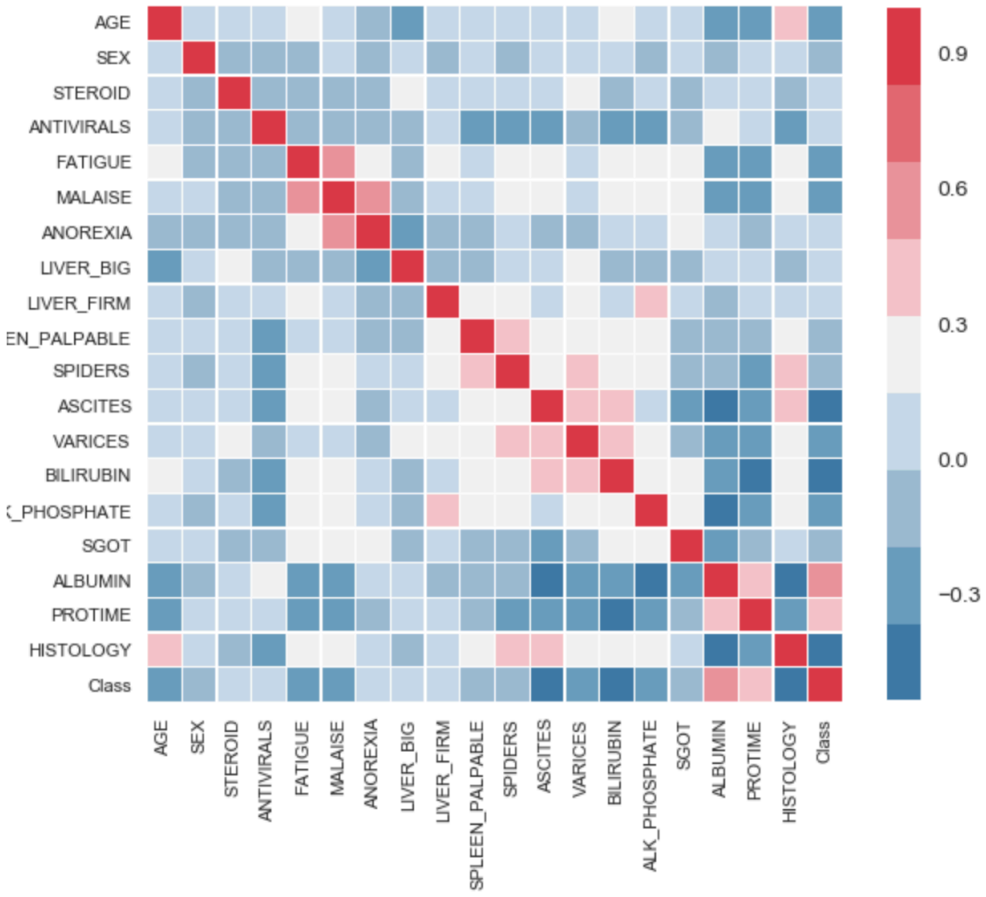


Figure 3.40: Correlation Analysis for all variables

By looking into this graph it can be conclude that there is no strong linear correlation between variables.

3.3.3 Attributes for dataset-Hepatitis

AGE	Age of the patient
GENDER	Gender of the patient
BMI	Body mass Index
FEVER	Fever
NAUSEA	Vomiting sensation
HEADACHE	Head ache
DIARRHOEA	Diarrhea
FATIGUE	Tiredness
EPIGASTRIC PAIN	Discomfort in upper abdomen
WBC	White blood cells
RBC	Red blood cells
HGB	Hemoglobin protein
PLATELETS	Platelet count
AST 1	Aspartate Aminotransferase 1
ALT 1	Alanine Aminotransferase 1
ALT 4	Alanine Aminotransferase 4
ALT 12	Alanine Aminotransferase 12
ALT 24	Alanine Aminotransferase 24
ALT 36	Alanine Aminotransferase 36
ALT 48	Alanine Aminotransferase 48
RNA BASE	Ribo nucleic acid Base
RNA 4	Ribo nucleic acid 4
RNA 12	Ribo nucleic acid 4

3.4 Feature Selection

3.4.1 Machine Learning Techniques:

a) PCA

Principal Component Analysis is an unsupervised, non-parametric statistical technique commonly used to reduce machine learning dimensionality. PCA helps to reduce attributes to a specified number of attributes.

In Hepatitis B, PCA is used to eliminate 5 attributes out of 20 which in turn helps to increase the efficiency of the prediction model. Few attributes may not impact the efficiency of prediction due to the size of the dataset.

In the case of Liver Damage Prediction PCA is not used to eliminate features because the number of attributes was 10 and all are equally important from domain study.

For Hepatitis PCA is used by varying number attributes limit, to fetch better results. Algorithms Like SVC which work with a lower dimension of a dataset are also applied.

b) Recursive Feature Elimination

The Recursive elimination function is primarily a retroactive predictor range. It starts by constructing a model for each predictor with the full set of predictors and calculating a significant value. The lowest predictor(s) are removed, the model is rebuilt and the value scores are again determined. The analyst typically determines the number of sub-sets to analyze and the scale of each sub-set. The sub-set size is therefore an RFE tuning parameter. The subset scale, which optimizes performance parameters, is used to select predictors based on the rankings of importance. The optimal subset is then used for the final mode preparation.

Since RFE recursively removes features and builds a model based on remaining attributes this is applied to Hepatitis C only as it was time- consuming. 17 attributes were selected out of 30 using RFE.

c) SMOTE Oversampling:

SMOTE generates synthetic samples from the minority classes. This gives us a synthetically balanced or nearly class-balanced training set. Which can be used to train Machine learning techniques.

SMOTE is applied for the Liver Damage dataset the dataset samples increased from 583 to 832. Which helped models to train better. For Hepatitis C SMOTE have been applied which increased from 1385 to 1448. For Hepatitis C SMOTE didn't increase the sample to the expected range. Variants of SMOTE techniques gave the same number of sample count. By applying SMOTE on the hepatitis B dataset led to the overfitting of the machine learning models. As the initial number of samples were low compared to the other two datasets

3.4.2 Domain Background

a) Hepatitis

One million people in India are at risk every year of contracting hepatitis. Abrupt onset of prodromal symptoms including fatigue, malaise, nausea, vomiting, anorexia, fever, and right upper quadrant pain. Dark urine, jaundice and pruritus(itching) develop in few days. Prodromal symptoms reduce as jaundice appears .On physical examination it shows jaundice and hepatomegaly(enlargement of liver). There may be splenomegaly(enlargement of spleen) and cervical lymphadenopathy(enlargement of lymph).Liver function tests where Serum aminotransferase elevated and ALT(SGPT) is more elevated than AST (SGOPT).

Bilirubin is elevated upto 30mg/dl .ALP is normal or slightly elevated. CRP, ESR and immunoglobulin are raised. The Active or Chronic Hepatitis B surface antigen test (HBsAg) is a blood test administered to screen for hepatitis B virus infection. It means that the individual has hepatitis B infection when found together with certain antibodies. If it is Acute infection then HBsAg, Anti HBcAg IgM are positive. If it is Chronic infection then HBsAg , Anti HbcAg , IgG are positive. If it is Previous infection then it is HBsAg is negative . Anti HBs is positive. IgG Anti HBc is

positive.(since the patient had previous infection antibodies are formed before). Presence of HBeAg indicates high infectivity. Liver function tests where Serum aminotransferase elevated and ALT(SGPT) is more elevated than AST (SGOPT).

Bilirubin is elevated up to 30mg/dl.ALP is normal or slightly elevated. CRP, ESR, and immunoglobulin are raised. Anti – HCV is positive.[5]

b) Liver

In the beginning of liver damage, exhaustion, weakness, and loss of weight can be felt. Patients may experience jaundice, gastrointestinal bleeding, abdominal swelling, and discomfort during further stages of the operation. Yellowing of the skin. People may experience pain in the abdomen. In Gastrointestinal there is bleeding, dark stubble from digested blood, abdominal fluid, nausea, excessive quantities of gas passing, vomiting blood, or water retention. The entire body there is weakness, appetite loss, or a lowered production of hormones. Treatment can help, but this condition can't be cured requires a medical diagnosis. Lab tests or imaging often required. Chronic can last for years or be lifelong.

4. REQUIREMENTS

4.1.1 Hardware Requirements

- System : Pentium IV 2.4 GHz.
- Hard Disk : 500 GB.
- Ram : 4 GB

4.1.2 Software Requirements

- Operating system : Windows XP / 7
- Coding Language : Python
- Technologies : Machine Learning
- IDE : Jupyter Notebook

4.2.1 User Requirements

- Functional requirements:
- To monitor the values of health parameters from time to time and know whether the values fall within the safe range.
- Easy to use with a simple and user-friendly UI. UI design -dashboard that displays the report.
- Reliability - The model should not fail to predict when values are passed to it.
- Performance - Get the results after inputting the parameters within 10 seconds.

4.2.2 Data Requirements

- Users need to get lab test done before using the model.
- Then users need to fill the form details from lab reports to get the result.

5. SYSTEM DESIGN

5.1 System Architecture

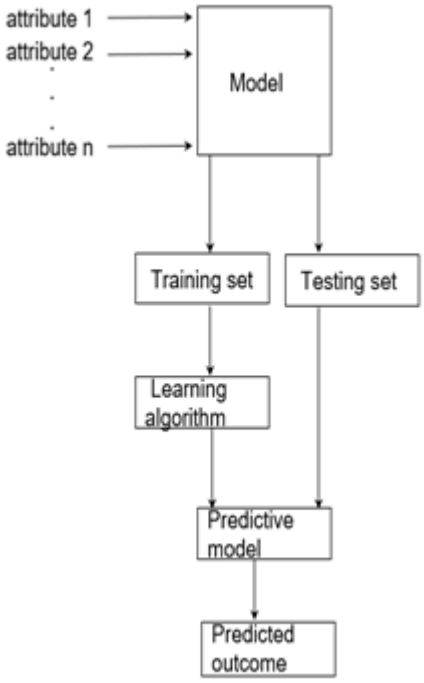


Fig 5.1 System Architecture

Dataset with n attributes are sent to the system. In the model data set is divided into training set and testing set. Training set is 70% and 30% is for testing accuracy. Then training set is analysed with machine learning algorithms SVM, Random forest, decision trees and gradient boosting. And both training and testing set is sent to predictive model to give the results.

For building the model, the training set is used. A hold-out dataset or test set is usually used to assess how well the model performs with knowledge outside of the training set.

5.1.1 Model:

The attributes required to predict Hepatitis or Liver Damage are loaded into the model.

5.1.2 Training Set:

From the model 70% of data is selected randomly from every attribute for Training Machine Learning Algorithms.

5.1.3 Testing Set:

From the model 30% of data is selected randomly from every attribute for testing Trained Machine Learning Algorithms.

5.1.4 Learning Algorithm:

The data fetched from Training set is used to train machine learning Algorithms.

5.1.5 Predictive Model:

This model takes trained algorithm from Learning Algorithm and Testing set. Here trained model predicts the outcome for each row in Testing set.

5.1.6 Predicted Outcome:

The predicted result of predictive Model is compared with the actual result of testing set. The accuracy of the Predictive Model is resulted.

5.2 Use case diagram

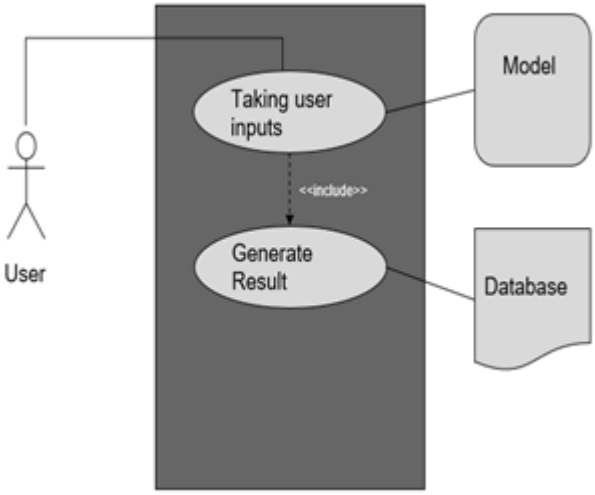


Fig 5.2 Use case Diagram

Actors:

- User: User must enter the lab report values into form and the predicted result is displayed.
- Model: User input is passed to predictive model which predicts the outcome. The outcome is displayed to User.
- Database: Consists of dataset of liver and hepatitis which are used to train predictive model.

Use cases:

- Taking user inputs: user inputs are taken the order of attributes in the input array matches the order of trained dataset input to the model when making a prediction.
- Generate Results: After predictive model is trained results are generated based on user inputs.

6. USER INTERFACE PRESENTATION

- **Liver UI**

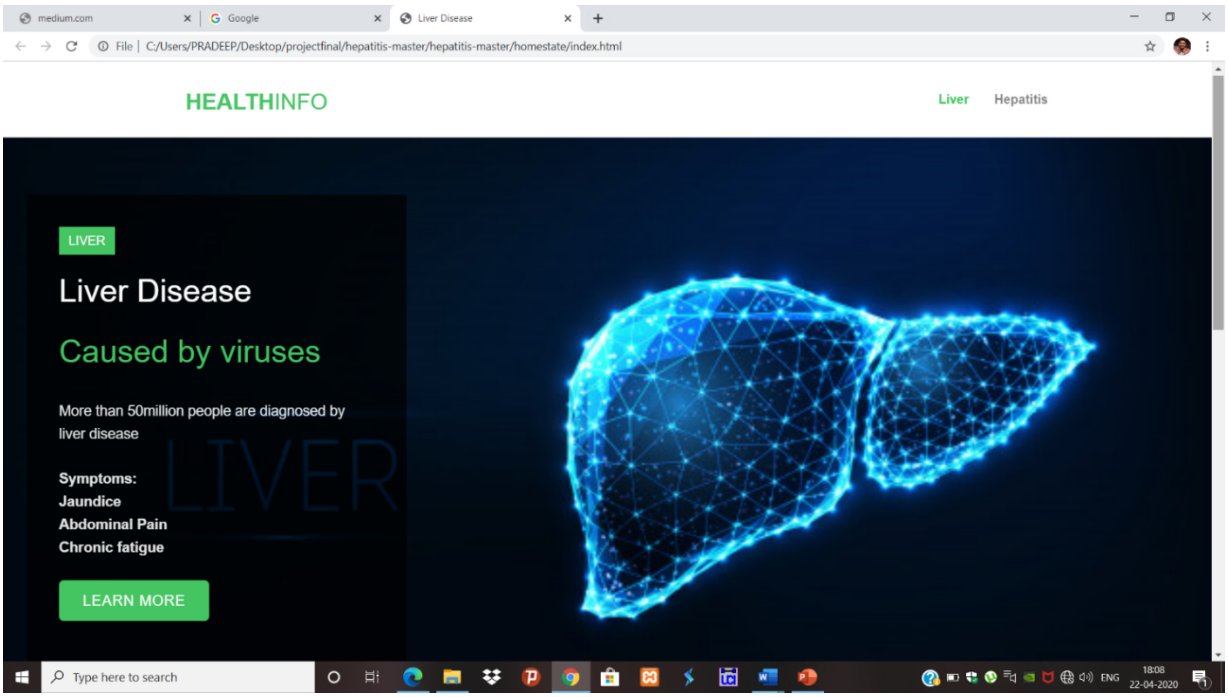


Fig 6.1 Liver UI

The above figure 6.1 represents the user interface of Liver form. It displays what causes the liver disease, how many people are diagnosed with liver disease and symptoms of liver disease. User can also learn more information about the liver disease.

6.1 Liver Form

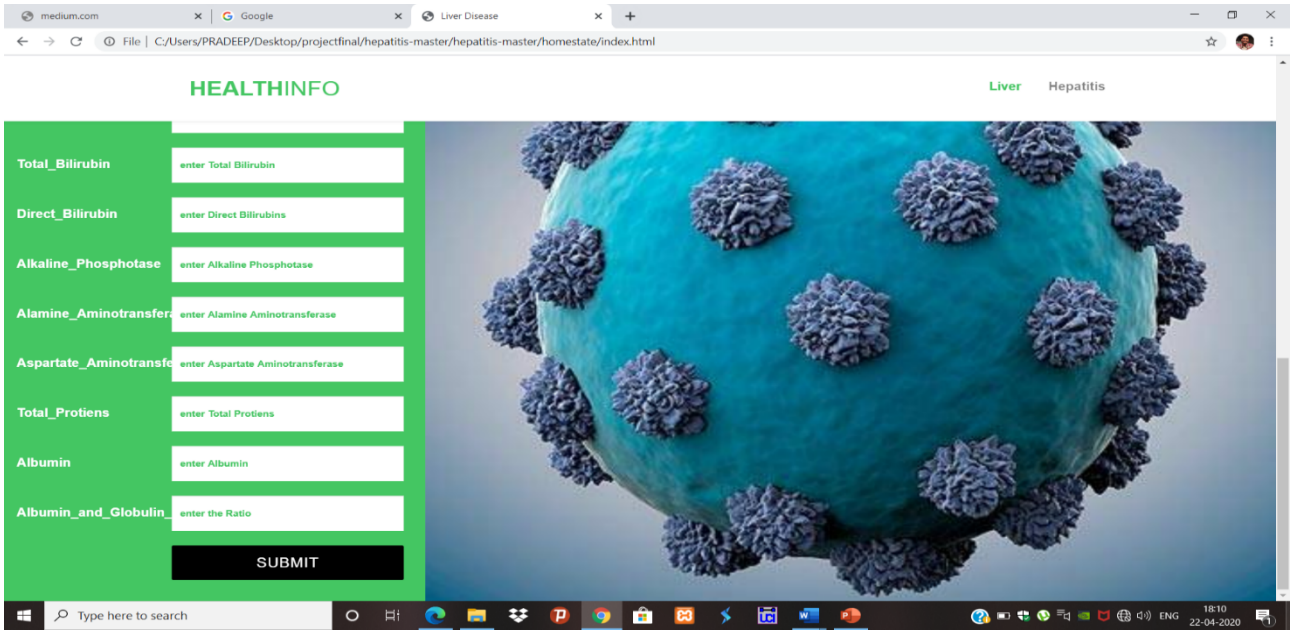


Fig 6.2 Liver form

Above Figure 6.2 represents the liver form. Where users fill in the laboratory results into the form. When the values are entered into the form it displays whether liver is damaged or no. UI is connected to the backend using Flask API where all the training and testing is done.

- **Hepatitis UI**

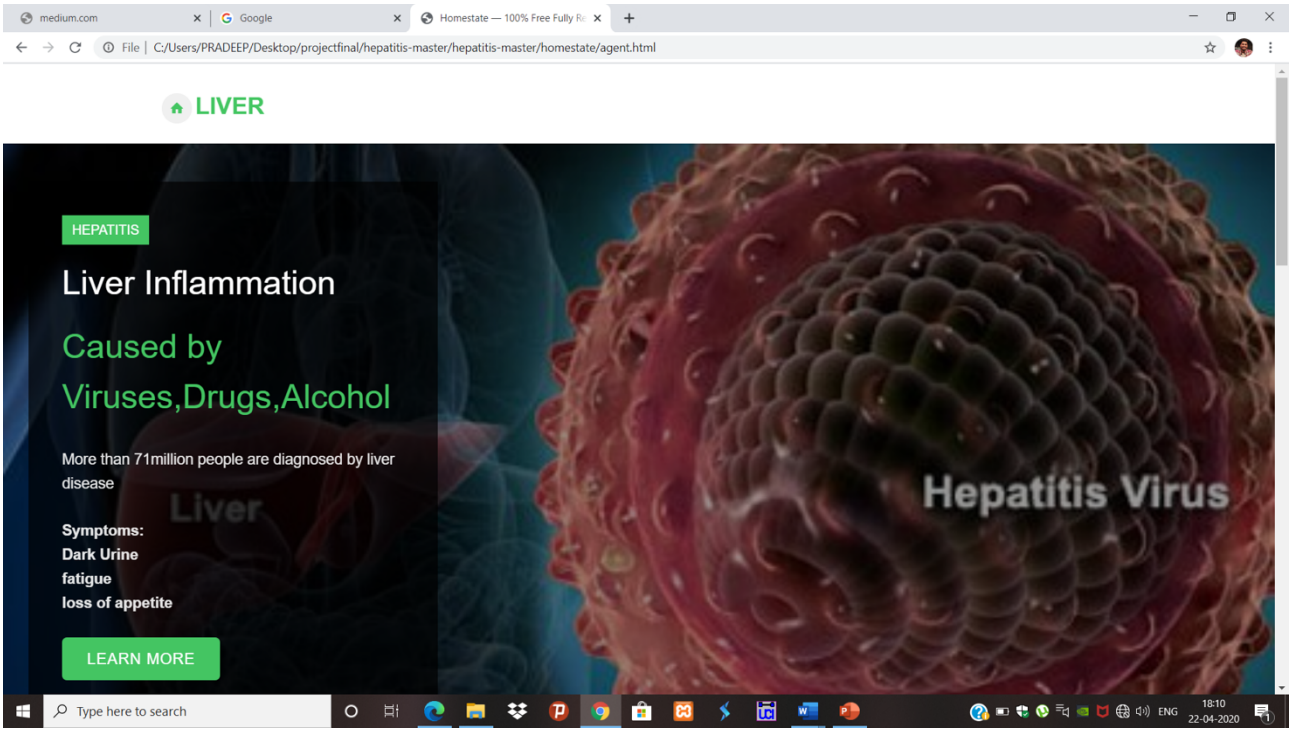


Fig 6.3 Hepatitis UI

The above figure 6.3 represents the user interface of Hepatitis form. It displays what causes the Hepatitis disease, how many people are diagnosed with Hepatitis disease and symptoms of Hepatitis disease. User can also learn more information about the Hepatitis disease.

6.2 Hepatitis Form

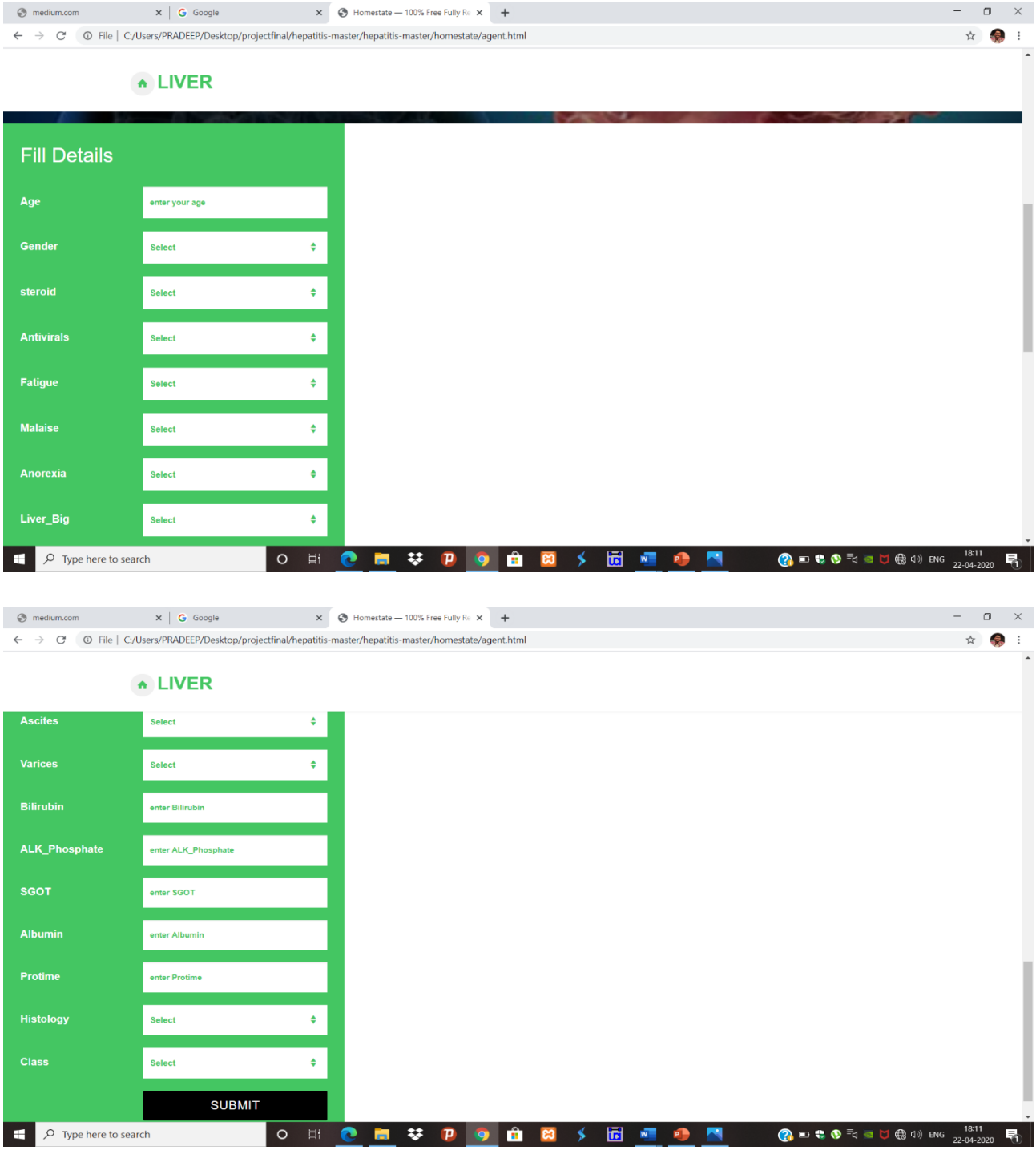


Fig 6.4 Hepatitis form

Above Figure 6.4 represents the Hepatitis form. Where users fill in the laboratory results into the form. When the values are entered into the form it displays whether liver person live or die. UI is connected to the backend using Flask API where all the training and testing is done.

7. RESULTS

7.1 Liver Damage Prediction Accuracy

LIVER	
Method	Accuracy
Logistic Regression	71.59
SVM	73.14
Random Forest	70.29
Smote: Logistic Regression	73.2
Smote: Random Forest	80.4

Liver damage dataset contains missing values so random forest algorithm has been implemented in those datasets. Presence of an enzyme or symptom can signify disease or damage presence. So this algorithm is implemented in Hepatitis B, Hepatitis C and Liver Damage predictions.

7.2 Hepatitis B Live or Die Prediction Accuracy

HEPATITIS B	
Method	Accuracy
Random Forest Classifier	74.19
PCA- 15 attributes, Random Forest	83.87
Gradient Boosting Classifier	68
Decision Tree	61.29

In Hepatitis B target column contains survival of patient. This dataset contains missing values so Random Forest algorithm has been implemented in those datasets. As there is a chance of overfitting in Decision tree the Oversampling technique SMOTE is not applied for Hepatitis B prediction. Since Hepatitis B contains less number of rows in the dataset, oversampling led to overfitting. Gradient Boosting Technique is used for Hepatitis B disease prediction after pre-processing. As this model is ensemble of weak prediction models and PCA is used to eliminate 5 attributes out of 20 which in turn helps to increase efficiency of prediction model. Few attributes may not impact efficiency of prediction due to size of the dataset.

7.3 Hepatitis C Fibrosis Stages Accuracy Prediction

HEPATITIS C – Fibrosis Stages	
Method	Accuracy
SVM	27.8
Random Forest	32.49
Logistic Regression	22.18
Decision Tree	24.18

HEPATITIS C – Fibrosis Stages after Pre-processing	
Method	Accuracy
SVM	27.44
Random Forest	20.58
Logistic Regression	22.18

HEPATITIS C – Enzyme and virus count	
Method	Accuracy
Logistic Regression (ALT)	75.45
Logistic Regression (RNA)	66.42
Decision Tree (ALT)	65.34
Decision Tree (RNA)	79.06

Hepatitis C dataset had no textual data and no missing values so initially SVM, Random forest, Logistic regression and decision tree were applied directly but did not give proper results. Later pre-processing has been done with applying PCA but accuracy did not improve significantly.

Finally weekly prediction has been done with considering key attributes ALT(Alanine Transferase), RNA(ribo nucleic acid). The prediction was whether these attributes show increasing signs during 24 - 36th week of prediction. Accuracy has been increased significantly when the above algorithms were applied.

8. CONCLUSION

1. **Hepatitis B:**

   Dataset of Hepatitis B contains text yes/no representing presence of an enzyme or symptom and values were missing in few attributes. So Imputer was used to replace missing values with mean value and yes/no values are converted into numerical value.

   After pre-processing Random forest, Gradient Boosting and Decision Tree algorithms were used to build models to predict hepatitis. PCA is used to fetch 15 attributes out of 20 attributes and above models are build. Which increased the prediction accuracy of Hepatitis B.

2. **Liver Damage:**

   There are several reasons for Liver damage it might be due to virus or lack of self-care. Dataset contains 10 attributes all of them found out to be necessary for prediction. Missing values are replaced by mean using imputer. To increase the size of the dataset SMOTE Over Sampling technique is used. Prediction models are trained before and after over sampling.

3. **Hepatitis C:**

   Hepatitis C dataset contained ALT and RNA enzyme and virus count which has been tested at regular weeks. As there is no missing values or textual values in dataset, machine learning models were directly applied. This gave us low accuracy.

   Due to this the attributes were categorized based on clinical ranges (e.g. age: 0-32 is grouped under single category). Now machine Learning models are applied, by doing this accuracy didn't change much. Then PCA is applied even after this accuracy didn't increase. It has been noticed that ALT/AST enzyme count are normal for hepatitis affected in one third of cases. So, ALT and AST columns have been dropped, prediction is done using same algorithms and PCA. By Using the Over Sampling technique SMOTE increased not more than 10% of the dataset. Even after doing this, the accuracy did not increase above 32.

   So having a lesser number of records for Hepatitis C blocked machine learning models to train efficiently, which outputted low accuracy.

9. FURTHER ENHANCEMENTS

The current model will perform prediction based on laboratory test dataset, more research will be done for larger dataset which increases accuracy. Further improvement is to take image dataset ie Ultrasound scanned images of the liver to enhance the prediction.

10. REFERENCES

[1] “Application of Machine Learning Classification Algorithms on Hepatitis Dataset”, International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 16 (2018) pp. 12732-12737

[2] “Prediction of Liver Fibrosis stages by Machine Learning model: A Decision Tree Approach”, Heba Ayeldeen∗1, 2, Olfat Shaker 1,3, Ghada Ayeldeen 1,3, Khaled M. Anwar 2 1 ISI Research Lab - www.isirlab.net

[3] “Liver Patient Classification Using Intelligent Techniques”, Anju Gulia et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (4) , 2014, 5110-5115

[4] “Liver Disease prediction by using different decision tree techniques”, International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.8, No.2, March 2018