Statistical Machine Learning Approaches to Liver Disease Prediction

A PROJECT REPORT

Submitted by

Abishek S K **711319CS002**

Anbu V **711319CS011**

Aswin Kumar H **711319CS019**

Jevanantham N **711319CS059**

Team ID : PNT2022TMID13793

TABLE OF CONTENTS

INTRODUCTION

1.1 Project Overview

In India, liver disease is the tenth leading cause of death, accounting for 2.95% of all deaths. According to the WHO, liver disease is one of the leading causes of death in India. It has become a major threat in India, with approximately 10 lakhs new cases diagnosed each year. One of the most important aspects of automated disease diagnosis and prediction are data mining. Data mining algorithms and techniques are used to analyses medical data. Liver disorders have skyrocketed in recent years, and liver disease is now in several countries, one of the leading causes of death. the patient datasets are evaluated for designing classification models to predict liver disease. In three phases, the above study used feature implementation and comparative analysis to improve accuracy rate of liver patients. The min max normalization algorithm is used on the existing liver patient datasets obtained from the Data sets in the first process. In the second stage of liver dataset prediction, PSO feature selection is used to obtain a subset (data) of liver patient dataset from whole normalized liver patient datasets that contains only significant attributes.

1.2 Purpose

To Detect disease, healthcare professionals need to collect samples from patients which can cost both time and money. Often, more than one test or many samples are needed from the patient to accumulate all the necessary information for a better diagnosis. The most routine tests are urinalysis, complete blood count (CBC), and comprehensive metabolic panel (CMP). These tests are generally less expensive and can still be very informatics. Liver-related disease accounts for 70 percent of deaths worldwide. There is a need to find better ways to detect and diagnose liver disease with more accuracy. Most importantly, tests of liver function need to be available and affordable to patients. To avoid the expensive and invasive tests, the application of statistical machine learning techniques to CMP results for the extraction of information for a clinician might be helpful for diagnosis.

Exploratory data analysis method are extremely important in healthcare; that can predict patterns across data sets to facilitate the determination of risk or diagnostic factor for disease with more speed and accuracy. The use of these methods can allow for earlier detection and potentially

prevent many cases of liver disease from progressing to the point of needing biopsy or complex treatment.

2. LITERATURE SURVEY

2.1 Existing Problem

It is the vital organ in the human body. It is most found in the upper right quadrant of the abdomen, beneath the diaphragm and above the ovaries.

The structure extends into the upper left quadrant to a small extent. It regulates the greater number of chemical levels in the blood and excretes a substance known as bile. It regulates the hemoglobin for use of the iron content. It produces the cholesterol and special proteins which aid in the transmit of fats throughout the body.

It filters the blood from the digestive tracts to filter blood from the digestive tract before it is distributed throughout the rest of the body. In moreover the liver detoxifies Chemicals and digests drugs. As it does so, the liver covers up bile that eventually ends up in the intestines. The liver also Generates proteins which are necessary for blood clotting and other functions. The liver is in control of many dangerous functions within the body, when it becomes ill or damaged, the loss of those Function can cause major harm to the body.

Kinds of liver disease:

There are some of the several diseases can cause liver damage.

i)Catching a viral infection, regularly drinking too much alcohol may cause the Hepatitis

ii)A gene that runs in families and may be passed from parents to children may cause the Haemochromatosis

iii)While there is a problem in the immune system it may be causes Primary biliary cirrhosis. The training dataset, which was created by P.Rajeswari,G Sophia Reena [2010] which is gathering information from the UCI repository, consists of 345 cases with 7 different properties. The data classification based on the algorithm is Naive Bayes algorithms, FT tree algorithms and K-Star algorithms. In the algorithm FT tree algorithm is used to test liver disease and faster and

more accurate when compared to other algorithm (98% of accuracy). Meanwhile experimental result says that classification accuracy is found good compared to other two algorithm [1].

The training dataset, which was created by Sa'diyah No or No-vita Alfisahrin [2013] which is being identified by the liver diseases by 10 different attributes by diseases person. The algorithm used behind those experiment is Decision tree, Naive Bayer algorithm, NB tree algorithm. The result shows that Naïve Bayer given more accuracy whereas NB tree algorithm gives fast computation time. On completion of research, we conclude that NB tree algorithm gives accuracy and perfect result compared to other two algorithm [2].

The physician required more clinical care of disorder of liver diseases by Dhamodharan S [2014]. He mainly undergoes the liver diseases like liver cancer, hepatitis, cirrhosis. His main objective the above-mentioned diseases are free for human. Naïve Bayer algorithm and FT tree algorithm are used by him during the research. The result says that Naïve algorithm gives more accuracy than FT tree algorithm [3].

Here we applied the data mining technique by S. E. Seker, Y. Unal, Z. Erdem, and H. Erdinc Kocer et al [2014]. He collected from 16,380 analysis of unique dataset. The algorithm used behind the experimental research is KNN, SVM, MLP or Decision tree algorithm. The analysis can be used for future scholar person and useful for detecting the anomaly on the analysis [4].

A.S. Annesh Kumar and Dr.C.JothiVenkateswaran [2015] has given brief explanation about category of liver diseases. He used the algorithm named as feature selection and fuzzy K-means classification. The same attribute value is used for many liver diseases to resolve the problem in liver. Fuzzy K-means classification gives better accuracy and expected output than feature selection. The accuracy is over 94 percent by using Fuzzy K-means classification [5].

Christopher N suggested a system for the diagnosis of medical diseases, taking into account six benchmarks: liver, hepatitis heart, diabetes, breast and lymph disorders. The researchers developed WSO and C4.5-based systems.

2.2 References

C Geetha, AR Arunachalam, EVALUATION BASED APPROACHES FOR LIVER DISEASE PREDICTION USING MACHINE LEARNING ALGORITHMS – Jan 2021.

Md Fazle, SM Mahedy Has an, Arifa, Md Asif, Md kamrul, PREDICTION OF LIVER DISORDERS USING ML ALGORITHMS-2020

Ajay S Singh, MD Irfan, Abishek, PREDICTION OF LIVER DISEASE USING CLASSIFICATION ALGORITHMS – 2018

Sanjay Kumar , Sarthar Katyal , EFFECTIVE ANALYSIS AND DIAGNOSIS OF LIVER DISORDER BY DATA MINING - 2018

Proposed Method

The system is used here for the machine learning that the model gets trained, then it got test. At last, the model will get efficient accuracy and will predict the result.

At initial stage, the main page that user is asked to enter the age, gender, total Bilirubin, direct Bilirubin, total proteins, albumin, A/G ratio, SGPT, SGOT and alkphos. By using those values, the we can find the blood test of the patient. After taking the above values from the user, the values are compared to the trained data which is trained by ourselves which gives the proper output and which is efficient one.

The system has following advantages:

1. No medical expertise required: The people who are need to check the liver diseases no need expertise who is using the liver prediction app by gives efficient result. The person is need to give only essential values like age, gender, albumin and then we will get result.
2. High accuracy: The system which get result will be cent percent result and we used data set for making this application. That's why the result might be highly efficient.
3. Immediate result: The user can found result immediately while entering the input values rather than waiting for an hour in hospital.

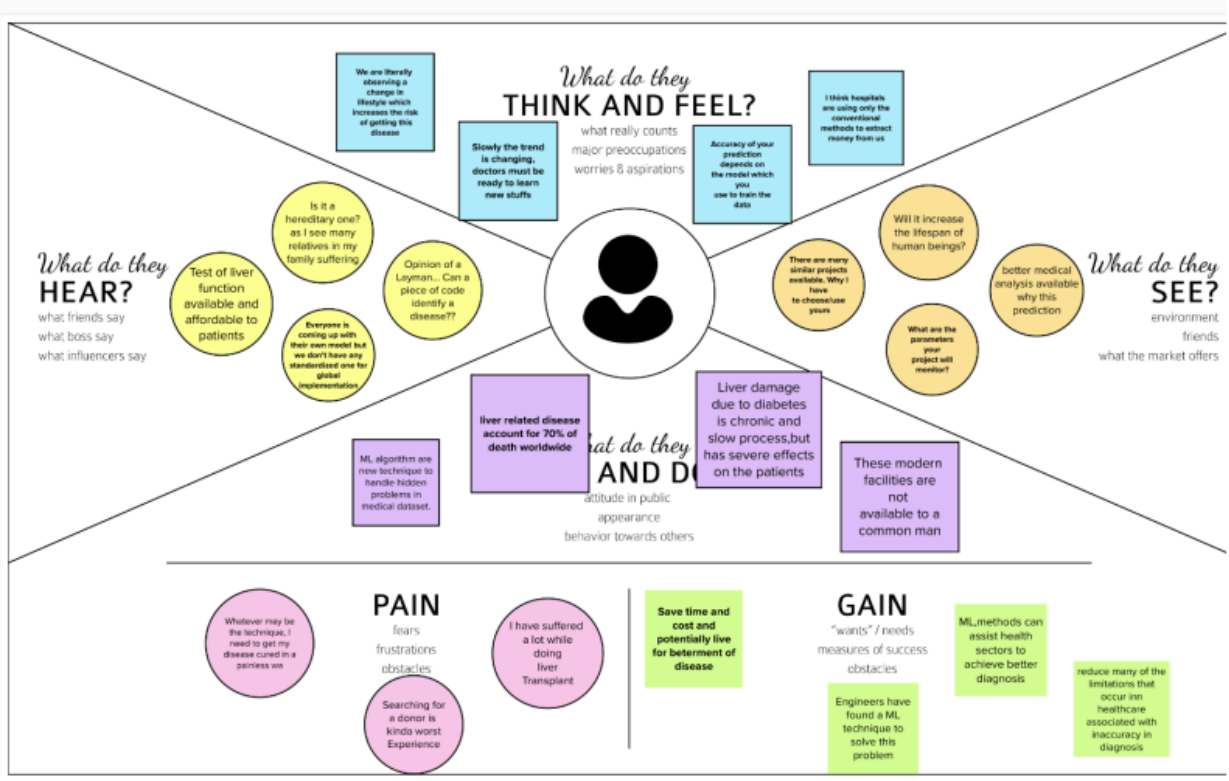
General workflow of the system related to creation and working:

The application mainly consists of the following tasks:

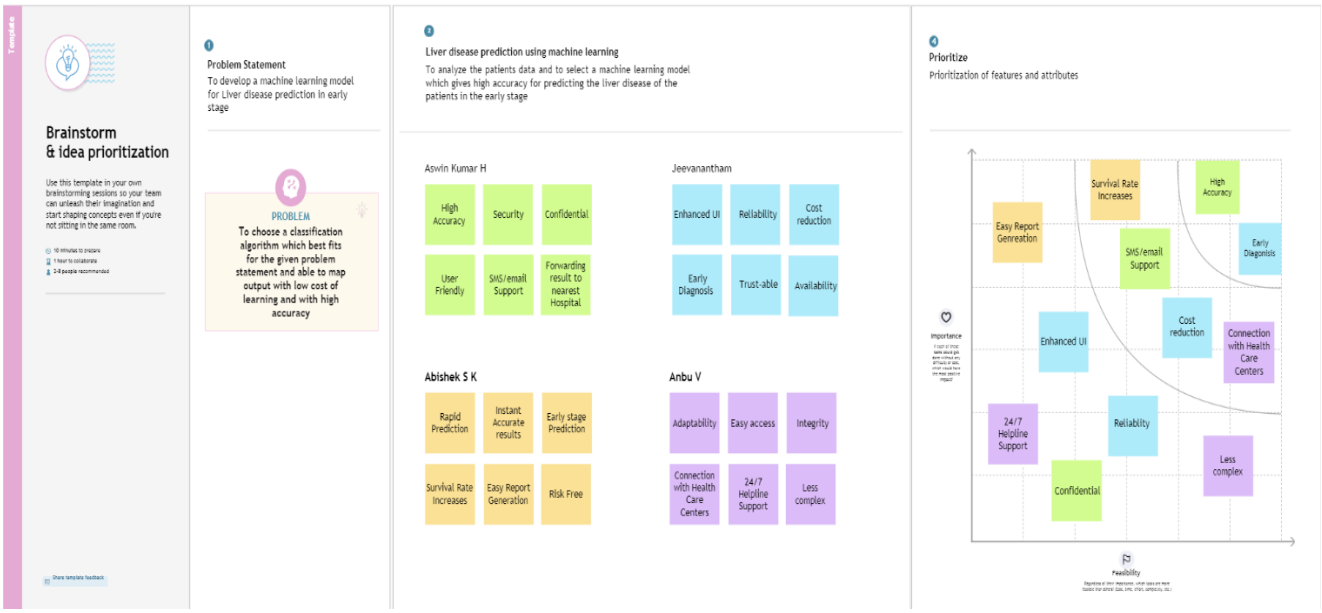
- Building and training the system: The project contains the frontend and backend and the database is used. The front will appear to the user while the backend cannot view by the user. The dataset will be trained and tested when it is initiated.
- Testing the models: In the time of phase the dataset will be tested and which is formed in previous data phase and will be more accurate.
- Entering details and prediction: In this phase, the backend user comes into play. The user which will enter the values like blood test using GUI of the app. The app will match the trained dataset and gives the efficient values. Then it will give the user has “RISK” or “NO RISK” as a output.

3.IDEATION & PROPOSED SOLUTION

3.1 Empathy Map Canvas



3.2 Ideation & Brainstorming



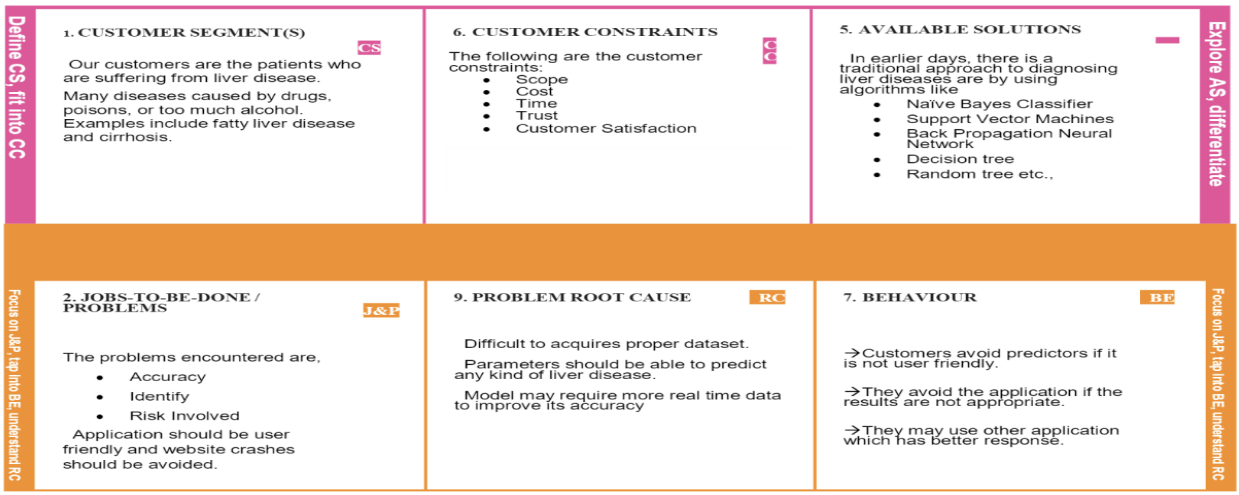
3.3 Proposed Solution

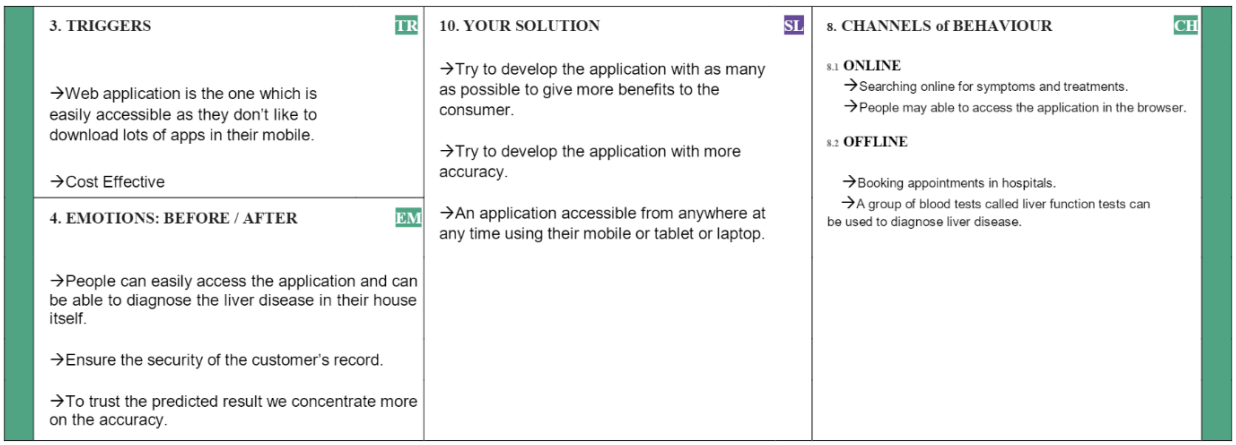
S.No	Parameter	Description
1.	Problem Statement (Problemto be solved)	Discovering the existence of liver diseases at early stage is a complex task for doctors. The challenge is to predict the liver disease patient fastand accurate and to diagnose the patients in early stage .
2.	Solution description	Machine learning model which usesstatistical data to predict the liver disease of the patients.

3.	Uniqueness	Accurately classifies the intensity of the liver disease from the patients concentrating on relationship between a key list of enzymes, proteins, age and gender using themto predict the likeliness of the liver disease
4.	Social Impact	• Capable of predicting the liverdisease in early stage • Works accurately and preciselyto predict the liver disease • Doctors can be able to diagnose the live patients in early stage to save many lives
5.	Business Model (RevenueModel)	• This system can be integrated with any Health sector domain,It solves the complex process of predicting the liver disease of patients and makes ease to the doctors to diagnose theliver disease. • The user can be able to getconsulting with doctors
6.	Scalability of the Solution	• Can be extended to predict

		many Classifications of diseasesin early stage • This can be integrated to withany hospitals and health sectors to get patient recordssecurely throughAPIs

3.4 Problem Solution Fit





4.REQUIREMENT ANALYSIS

4.1 Functional Requirements

FR No.	Functional Requirement(Epic)	Sub Requirement (Story / Sub-Task)
FR-1	**User Registration**	Registration through Form Registration through Gmail
FR-2	**User Confirmation**	Confirmation via EmailConfirmation via OTP
FR-3	**Website Entry**	Collecting user's data and storing it in theDatabase
FR-4	**Permissions**	Location, Storage, Contacts

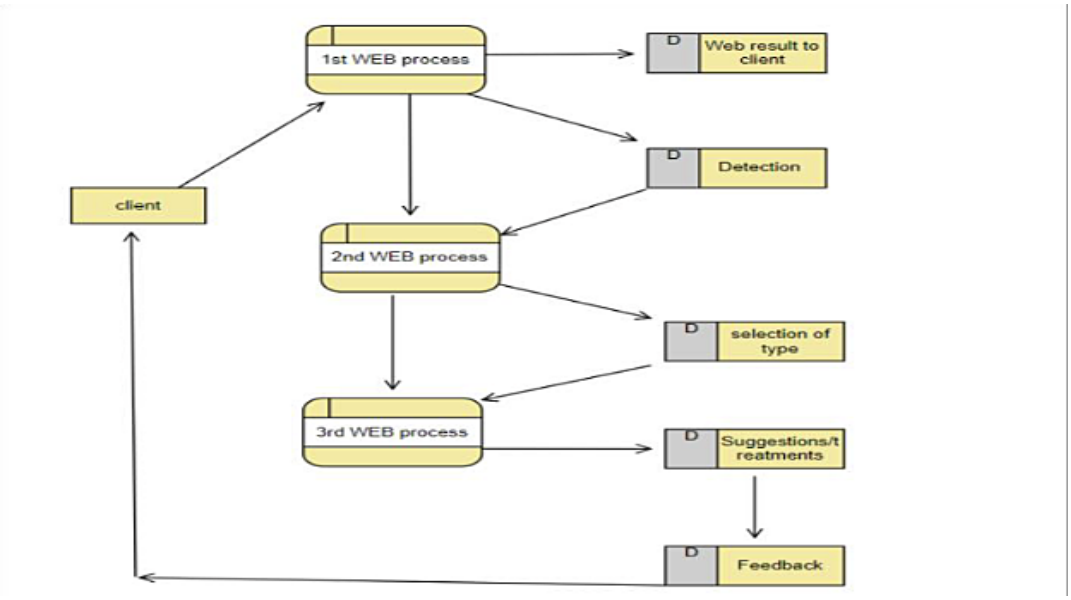
4.2 Non-Functional Requirements

FR No.	Non-Functional Requirement	Description
NFR-1	**Usability**	Defines how difficult it will be for a user to learn Andoperate the system. Usability can be accessed from different points.
NFR-2	**Security**	Security requirements ensure that the software is protected from unauthorized access to the system and its stored in data.
NFR-3	**Reliability**	Reliability defines how likely it is for the software to work without failure for a given period. Reliability decreases because of bugs in the code , hardware , failures and problems with other system component.

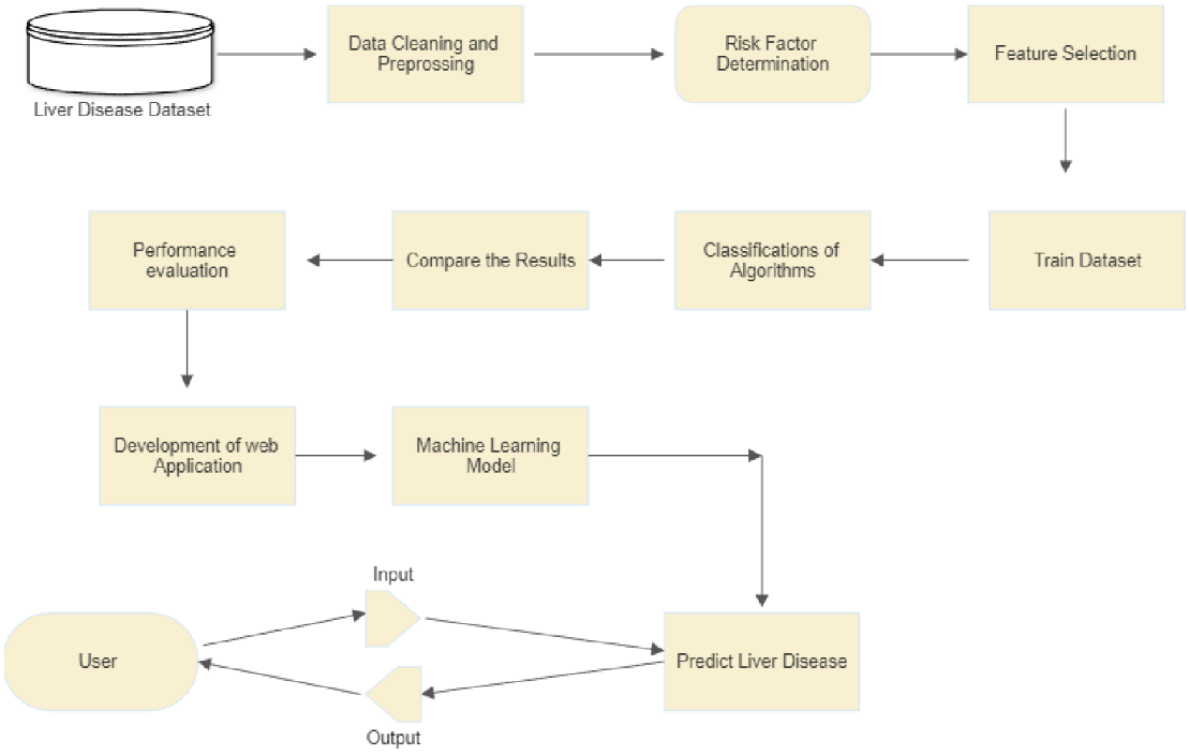
NFR-4	**Performance**	It is quality attribute that describers responsivenessof system to the various user interactions with it.
NFR-5	**Availability**	It is gauged by period that system's functionality &services are available for use with all operations.
NFR-6	**Scalability**	Scalability describes how the system must grow without negative influence on its performance. This means serving more users , processing more data , doing more transactions . Website traffic limit must be scalable enough to support 2,00,000 users at a time.

5.PROJECT DESIGN

5.1 Data Flow Diagram



5.2 Solution & Technical Architecture



5.3 User Stories

User Type	Functional Requirement(Epic)	User Story Number	User Story / Task	Acceptance criteria	Priority	Release
Customer (Web User)	Login	USN-1	As a user,I can login for the application by entering my email	I can access my account / dashboard	High	Sprint-1
		USN-2	As a user,I will login,confirmation email once I have registered	I can receive confirmation email	High	Sprint-1
		USN-3	As a user,I can login for the application through Mobile number	I can login and access the application	Medium	Sprint-2

	Dashboard	USN-4	As a user,I need to enter my Details	I can get information as per details	High	Sprint-3
	Dashboard	USN-5	As a user,I need to enter my Test Details	I can get result based on test details	High	Sprint-3
Administrator	Services	USN-6	As a admin I need to provide valid result	I can get a result	High	Sprint-3
		USN-7	As a adminI need to provide valid/useful Suggestions	I can get suggestions	Medium	Sprint-3
	Mass Data Process	USN-8	As a admin need to collect all the details and information	I can use it for later period	High	Sprint-4
		USN-9	As a adminI need to store all the details and information	I can use it for later period	High	Sprint-4
Hospital Administrator	Login	USN-10	As a adminI need to login and access details of customers	I can use for it further next step process	High	Sprint-4
	Dashboard	USN-11	As a adminI need to proceed the details with case head	I can use for further next step process	High	Sprint-4

6.PROJECT PLANNING & SCHEDULING

6.1Sprint Planning & Estimation

Sprint	Functional Requirement(Epic)	User Story Number	User Story/ Task	Story Points	Priority	Team Members
Sprint-1	Registration	USN-1	As a user,I can register for the application by entering my email, password, and confirming my password.	5	High	Aswin Kumar H

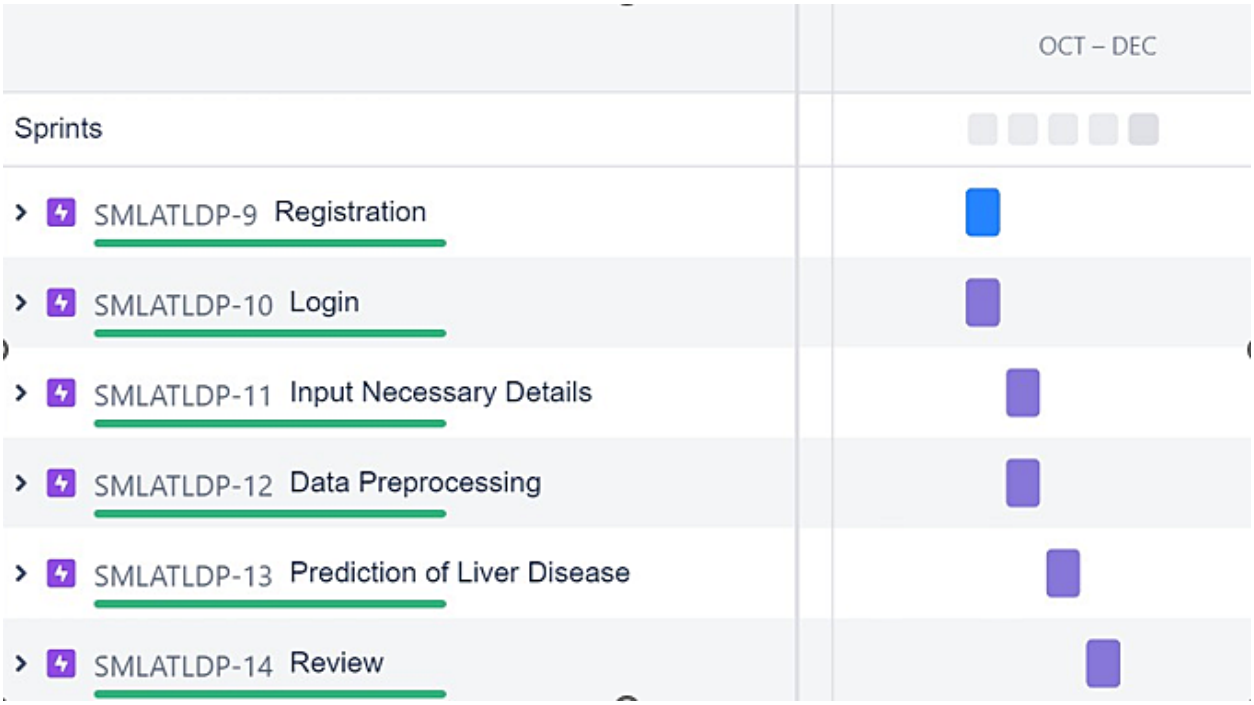
Sprint-1		USN-2	As a user, I will receive confirmation emailonceI have registered for the application	5	High	Aswin Kumar H
Sprint-1	Login	USN-3	As a user,I can log into the application by entering email& password	10	High	Jeevananth am N
Sprint-2	Input Necessa ryDetails	USN-4	As a user, I can give InputDetails to PredictLikeliness of Liver Disease.	15	High	Jeevananth am N
Sprint-2	Data pre-processing	USN-5	Transform raw data into suitableformat for prediction.	5	High	Abishek S K
Sprint-3	Prediction ofLiverDisease	USN-6	As a user, I can predict LiverDiseaseusing machine learning model.	15	High	Abishek S K
Sprint-3		USN-7	As a user, I can get accurate prediction of liver disease.	5	Medium	Anbu V
Sprint-4	Review	USN-8	As a user, I can give feedback of the application.	20	High	Anbu V

6.2 Sprint Delivery Schedule

Sprint	**Total StoryPoints**	**Duration**	**Sprint StartDate**	**Sprint End Date(Planned)**	**Story PointsCompleted (as on Planned End Date)**	**Sprint Release Date(Actual)**
Sprint-1	20	6 Days	24 Oct 2022	29 Oct 2022	18	29 Oct 2022
Sprint-2	20	6 Days	31 Oct 2022	05 Nov 2022	17	05 Nov 2022
Sprint-3	20	6 Days	07 Nov 2022	12 Nov 2022	18	12 Nov 2022
Sprin	20	6 Days	14 Nov	19 Nov 2022	17	19 Nov 2022

t-4			2022			

6.3 Reports from JIIRA



7.CODING & SOLUTIONING

7.1 Feature 1

index.html

```
<!DOCTYPE html>
<html>
<head>
  <title>Liver Disease Prediction</title>
```

```
<meta name="viewport" content="width=device-width, initial-scale=1">
<style>
body {
  font-family: Arial, Helvetica, sans-serif;
  background-image:url("../static/images/bimg.png");
  background-repeat: no-repeat;
  background-attachment: fixed;
  background-size: cover;
}

.navbar {
  overflow: hidden;
  background-color: rgba(252, 17, 107, 0.74);
}

.navbar a {
  float: left;
  font-size: 16px;
  color: white;
  text-align: center;
  padding: 14px 16px;
  text-decoration: none;
}

.dropdown {
  float: left;
  overflow: hidden;
}

.dropdown .dropbtn {
  font-size: 16px;
  border: none;
  outline: none;
  color: white;
  padding: 14px 16px;
  background-color: inherit;
  font-family: inherit;
  margin: 0;
}

.navbar a:hover, .dropdown:hover .dropbtn {
  background-color:black;
}
```

```
.dropdown-content {
  display: none;
  position: absolute;
  background-color: #f9f9f9;
  min-width: 160px;
  box-shadow: 0px 8px 16px 0px rgba(0,0,0,0.2);
  z-index: 1;
}

.dropdown-content a {
  float: none;
  color: black;
  padding: 12px 16px;
  text-decoration: none;
  display: block;
  text-align: left;
}

.dropdown-content a:hover {
  background-color: #ddd;
}

.dropdown:hover .dropdown-content {
  display: block;
}
input[type=text], select {
  width: 100%;
  padding: 12px 20px;
  margin: 8px 0;
  display: inline-block;
  border: 1px solid #ccc;
  border-radius: 4px;
  box-sizing: border-box;
  border-radius: 20px;
}
.registerbtn {
  background-color: #4CAF50;
  color: white;
  padding: 16px 20px;
  margin: 8px 0;
  border: none;
  cursor: pointer;
```

```
  width: 16%;
  opacity: 0.9;
  border-radius: 15px;

}

.registerbtn:hover {
  opacity:1;
}



input[type=submit]:hover {
  background-color: #45a049;
}
.center {
  /* margin: auto; */
  margin-left: 20px;
  width: 40%;
  border: 2px solid black;
  padding: 10px;
  border-radius: 25px;
}

</style>
</head>
<body>
    <!-- <p style="background-image: url('bimg.jpeg');"></p> -->

  <h2><p style="text-align:center; color:white;font-size: 50px">Liver Disease Prediction</p></h2>
  <div class="center">
  <form action="/predict"method="post">
    <input type="text" id="Age" placeholder="Age " name="Age" required="required">
    <br>
    <input type="text" id="Gender" placeholder="Gender (Male:1 , female:0)" name="Gender"
required="required">
    <br>
    <input type="text" id="Total_Bilirubin" placeholder="Total_Bilirubin" name="Total_Bilirubin"
required="required">
    <br>
    <input type="text" id="Direct_Bilirubin" placeholder="Direct Bilirubin" name="Direct_Bilirubin"
required="required">
    <br>
```

```
   <input type="text" id="Alkaline_Phosphotase" placeholder="Alkaline Phosphotase"
name="Alkaline_Phosphotase"required="required">
   <br>
   <input type="text" id="Alamine_Aminotransferase" placeholder="Alamine Aminotransferase"
name="Alamine_Aminotransferase" required="required">
   <br>
   <input type="text" id="Aspartate_Aminotransferase" placeholder="Aspartate_Aminotransferase"
name="Aspartate_Aminotransferase" required="required">
   <br>
   <input type="text" id="Total_Protiens" placeholder="Total Protiens" name="Total_Protiens"
required="required">
   <br>
   <input type="text" id="Albumin" placeholder="Albumin " name="Albumin" required="required">
   <br>
   <input type="text" id="Albumin_and_Globulin_Ratio" placeholder="Albumin and Globulin_Ratio"
name="Albumin_and_Globulin_Ratio" required="required">
   <br>
   <br>

   <button type="submit" class="registerbtn"><b>Predict</b></button>
</form>

</div>


</body>
</html>
```

negative_result.html

```
<!DOCTYPE html>
<html lang="en">
<head>
   <meta charset="UTF-8">
   <meta http-equiv="X-UA-Compatible" content="IE=edge">
   <meta name="viewport" content="width=device-width, initial-scale=1.0">
   <title>Result</title>
</head>
<body>
   <div style="text-align:center">
      <h1 style="text-align: center;">Result</h1>
      <h2>you are  safe</h2>
```

```
        <img src="/static/images/n.webp" style="height: 700px;width: 600px;"><br>
        <a href="/">Back to Home Page</a>

    </div>
</body>
</html>
```

positive_result.html

```
<!DOCTYPE html>
<html lang="en">
<head>
    <meta charset="UTF-8">
    <meta http-equiv="X-UA-Compatible" content="IE=edge">
    <meta name="viewport" content="width=device-width, initial-scale=1.0">
    <title>Result</title>
</head>
<body>
    <div style="text-align:center">
        <h1 style="text-align: center;">Result</h1>
        <h2>Chances of having Liver Disease is more, please consult a Doctor !!!
        </h2>
        <img src="/static/images/p.png" style="height: 600px;width: 600px;"><br>
        <a href="/">Back to Home Page</a>

    </div>
</body>
</html>
```

7.2 Feature 2

```
from flask import Flask, render_template, request, redirect, url_for, flash
import pickle

app = Flask(__name__)
filename="Model"
infile = open(filename,'rb')
model = pickle.load(infile)

@app.route("/")
def home1():
    return render_template("index.html")

@app.route("/predict",methods=["POST"])
def predict():
    Age=int(request.form['Age'])
    Gender=int(request.form['Gender'])
    Total_Bilirubin=float(request.form['Total_Bilirubin'])
    Direct_Bilirubin=float(request.form['Direct_Bilirubin'])
    Alkaline_Phosphotase=int(request.form['Alkaline_Phosphotase'])
    Alamine_Aminotransferase=int(request.form['Alamine_Aminotransferase'])
    Aspartate_Aminotransferase=int(request.form['Aspartate_Aminotransferase'])
    Total_Protiens=float(request.form['Total_Protiens'])
    Albumin=float(request.form['Albumin'])
    Albumin_and_Globulin_Ratio=float(request.form['Albumin_and_Globulin_Ratio'])

pre=model.predict([[Age,Gender,Total_Bilirubin,Direct_Bilirubin,Alkaline_Phosphotase,Alamin
e_Aminotransferase,Aspartate_Aminotransferase,Total_Protiens,Albumin,Albumin_and_Globuli
n_Ratio]])
```

```
    pre=str(pre[0])
    if pre=="1":
        return render_template("positive_result.html")
    else:
        return render_template("negative_result.html")



if __name__=="__main__":
    app.run(debug=True)
```

7.3 Feature 3

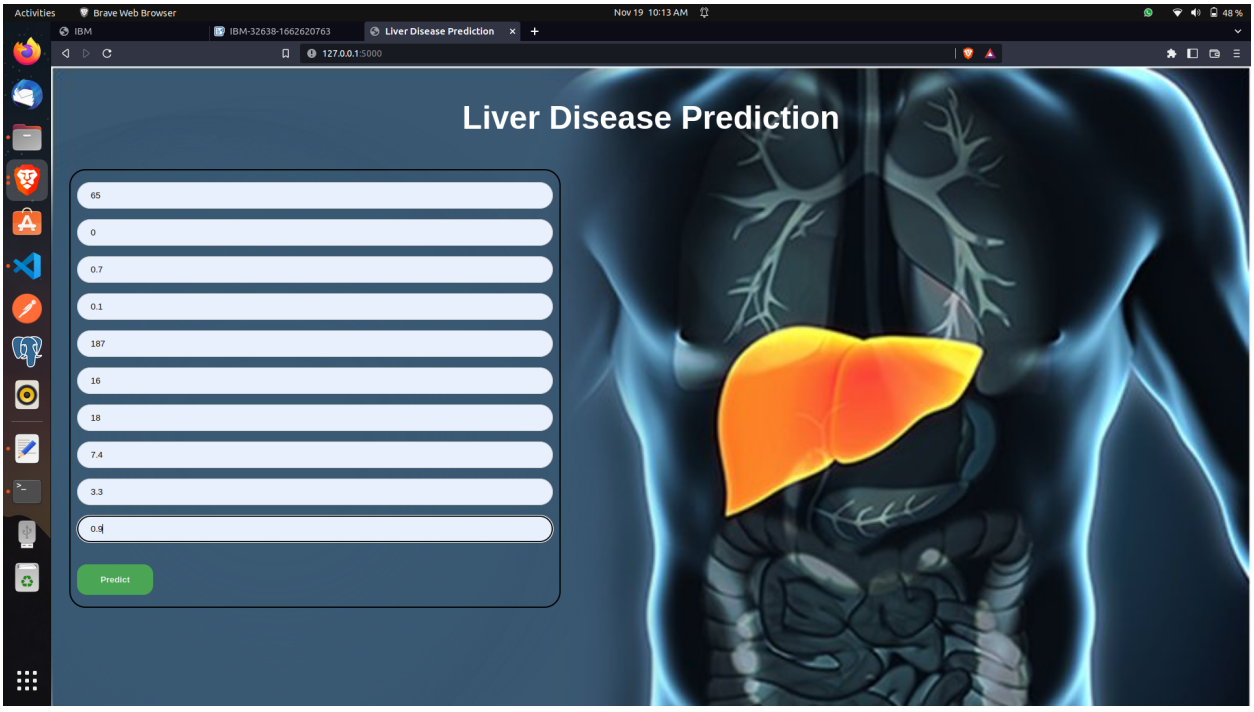
7.3.1 JUPYTER NOTEBOOK

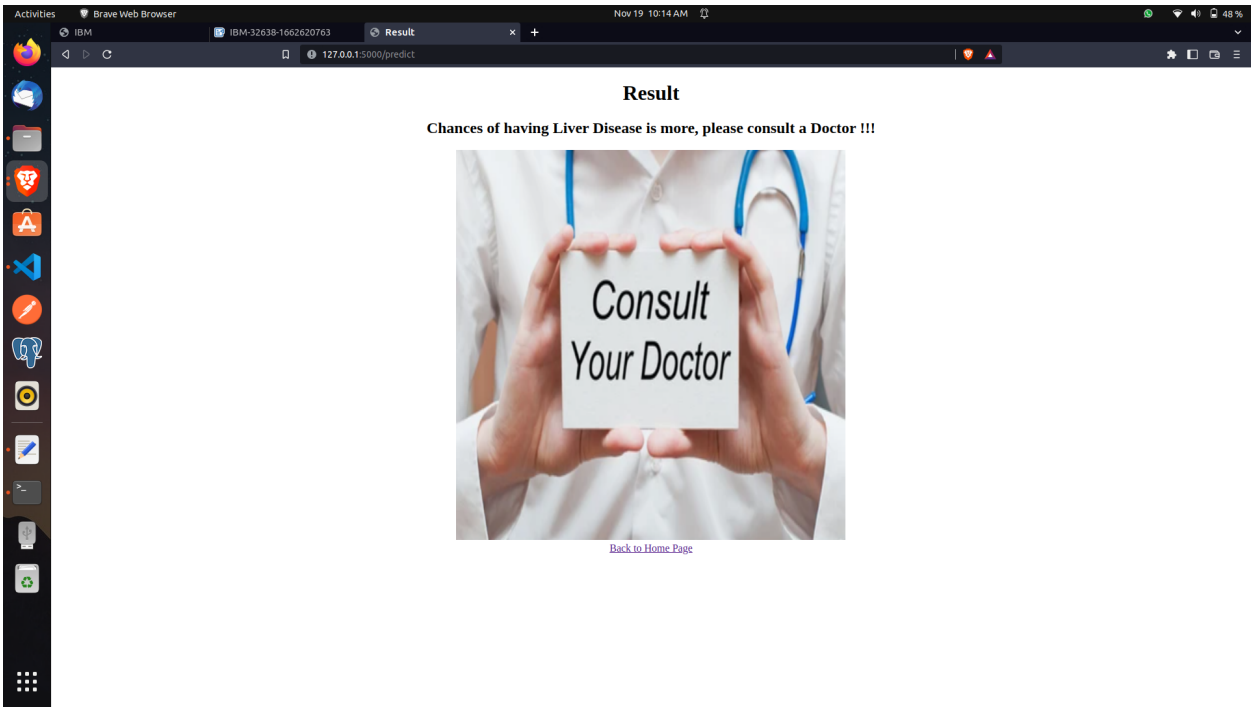
7.4 Database Schema

8.TESTING

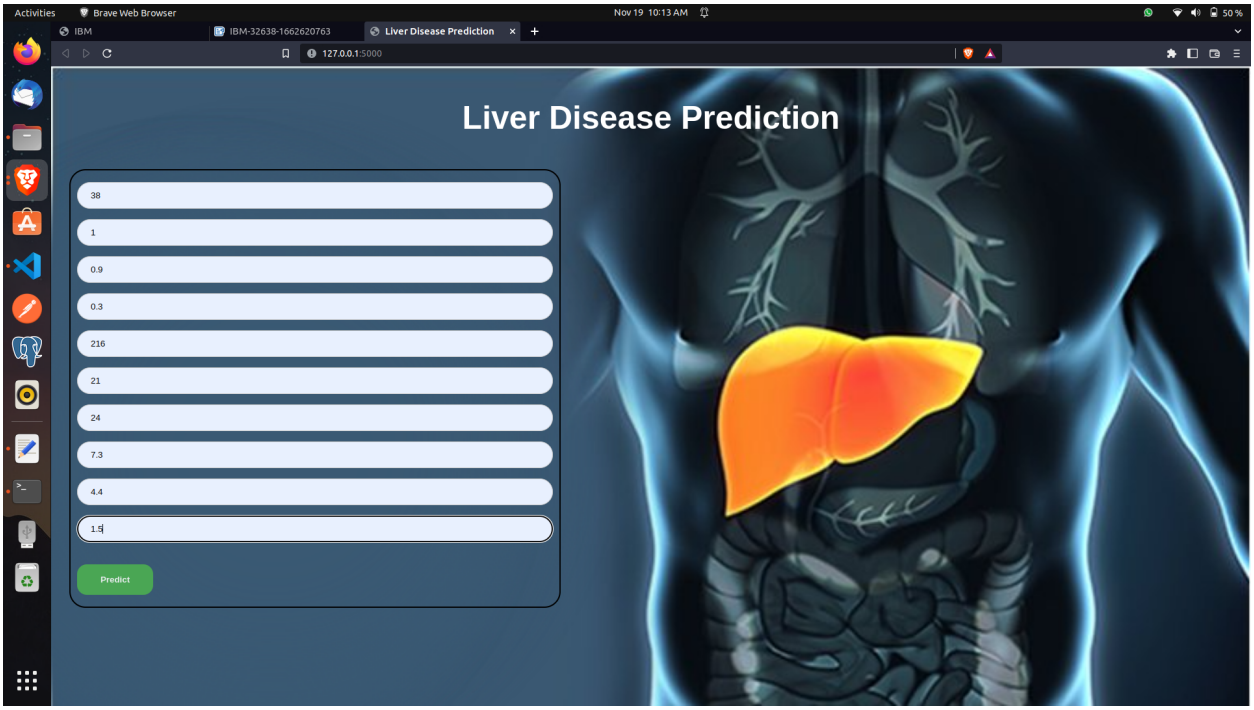
8.1 Test Cases &User Acceptance Testing

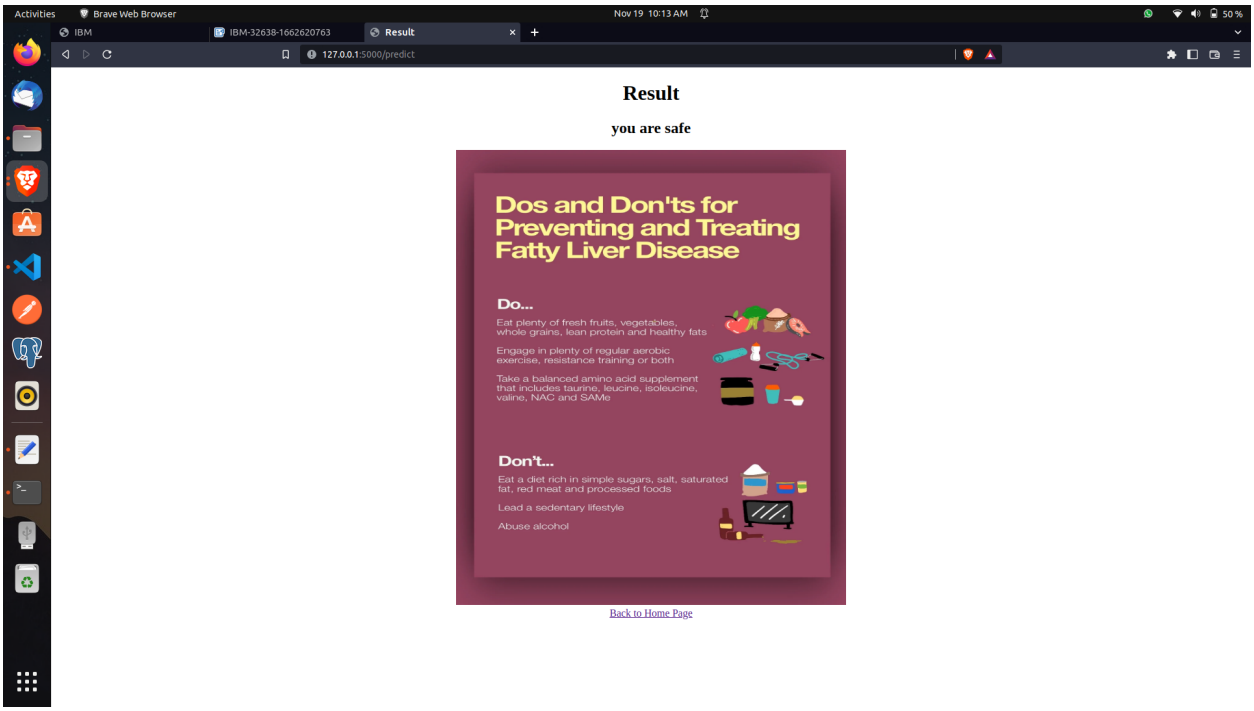
POSITIVE RESULT





NEGATIVE RESULT

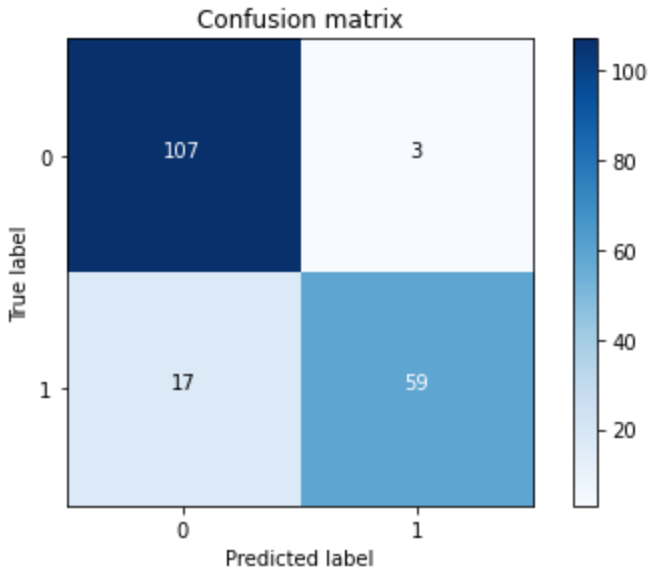




9.RESULTS

9.1Performance Metrics

```
Accuracy score for Decision Tree :  0.8924731182795699
```



	precision	recall	f1-score	support
0	0.86	0.97	0.91	110
1	0.95	0.78	0.86	76
accuracy			0.89	186
macro avg	0.91	0.87	0.88	186
weighted avg	0.90	0.89	0.89	186

10.ADVANTAGES & DISADVANTAGES

Advantages:

General

- Diagnoses, Grades and Stages:
  - Hepatitis C
  - Hepatitis B
  - Steatohepatitis
- Evaluates Abnormal Liver Function Tests
- Identifies Hepatotoxicity

- Clarifies uncertain diagnoses
- Confirms etiology of Liver Masses
- Defines extent of Necroinflammatory Activity
- Differentiates Fibrosis from Cirrhosis

Liver Transplant

- Identifies Acute Cellular Rejection
- Defines recurrence of Original Disease
- Identifies Progressive Fibrosis
- Diagnoses other Liver Processes

Disadvantages:

General

- Invasive
- Accessibility to the procedure
- Need for Training
- Repeated testing
- Cost

Sample

- Sampling error
- Intraobserver and Interobserver variations in interpretation
- Specimen Length and Width

11.CONCLUSION & FUTURE SCOPE

Conditions related to liver and heart are getting more and more common with time. With nonstop technological advancements, these are only going to increase in the future. Although people are getting more conscious of health Currently and are joining yoga classes, cotillion classes; still the sedentary life and luxuries that are continuously being introduced and enhanced; the problem is going to last long. So, in such a script, our design will be extremely helpful to the society.

With the dataset that we used for this design, we got 94 percentage for decision tree model, and though it might be delicate to get similar rigor with veritably large datasets, from this design's results, one can easily conclude that we can prognosticate the threat of liver conditions with percentage of 90 or more. moment nearly everybody above the age of 12 times has smartphones with them, and so we can incorporate these Results into an android app or iOS app. Also, it can be incorporated into a website and these app and website will be largely salutary for a large section of society.

1. APPENDIX

Source Code

- **Source code**

Github & Project Demo Link

- **Github**
- **Project Demo Link**