PROJECT DOCUMENTATION

INTRODUCTION

TOPIC: STATISTICAL MACHINE LEARNING APPROACHES

TO LIVER DISEASE PREDICTION

The improvement of patient care, research, and policy is significantly impacted by medical diagnoses. Medical practitioners employ a variety of pathological techniques to make diagnoses based on medical records and the conditions of the patients. Disease identification has been significantly enhanced by the application of artificial intelligence and machine learning in conjunction with clinical data. Data driven, machine learning (ML) techniques can be used to test current approaches and support researchers in potentially innovative judgments. The goal of this work was to use ML algorithms to derive meaningful predictors of liver disease.

1.1 PROJECT OVERVIEW

The number of patients with liver disease has been steadily rising as a result of excessive alcohol use, exposure to hazardous gases, ingestion of tainted foods such pickles and cucumbers, and drug usage. In an effort to lighten the load on doctors, this dataset was used to assess prediction systems. The data set consists of the patient&#39;s age, gender, and total bilirubin. Direct bilirubin, alkaline phosphatase, alamine aminotransferase, aspartate aminotransferase, total proteins, albumin, and the ratio of albumin to globulin.

1.1 PURPOSE

The purpose of the project is to help the patients to identify whether they are affected by liver disease by entering some medical data into the website which is easy to access and user friendly.

LITERATURE SURVEY

2.1 EXISTITING PROBLEM

We use Machine Learning approaches to the application of statistical machine learning techniques to results for the extraction of information for a clinician might be helpful for diagnosis. Exploratory data analysis methods are extremely important in healthcare, they can predict patterns across data sets to facilitate the determination of risk or diagnostic factors 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.

2.2 REFERENCES

Paper 1:
Software-based Prediction of Liver Disease with Feature Selection and Classification Techniques. [Jagdeep-Singh 1970–1980]
Today, everyone&#39;s health is a very essential concern, so it is necessary to offer medical services that are freely accessible to everyone. The primary goal of this study is to forecast liver illness using a software engineering methodology that makes use of feature selection and classification techniques. The Indian Liver Patient Dataset (ILPD) from the University of California, Irvine database is used to carry out the proposed research. The many variables of the liver patient dataset, including age, direct bilirubin, gender, total bilirubin, Alkphos, sgpt, albumin, globulin ratio, and sgot, among others, are used to forecast the risk level of liver illnesses.
On the Liver Patient dataset, several classification techniques are applied to determine accuracy, including Logistic Regression, Sequential Minimal Optimization, and K-Nearest Neighbor.

Biochemical Evaluation of Patients of Alcoholic Liver Disease and Nonalcoholic Liver Disease.[PRASAD.P.TORKADI 1979–1983]
The fundamental drawback of this approach is that, while the KNN algorithm predicts the outcome with a moderate degree of accuracy, it classifies the data according to the dataset&#39;s majority. Alcohol abuse over an extended period of time causes alcoholic liver disease (ALD). Because ALD patients are managed differently than individuals without ALD, accurate diagnosis is crucial. This system&#39;s objectives were to (1) compare the biochemical parameters of ALD and non-ALD patients to controls, and (2) determine whether these parameters can distinguish between ALD and non-ALD. The study involved 35 patients with acute viral hepatitis and 50 patients with alcoholic liver disease (ALD) in groups I and II, respectively. Our research shows that serum AST/ALT ratio, GGT, and ALP measurements may reliably distinguish ALD patients from NASH and acute viral hepatitis.

Paper 3:
Liver Disease Prediction using Naïve Bayes Algorithms. [Dr. S. Vijayarani 1816–1820]
Data mining has recently improved the simplicity of use for disease prediction in the healthcare sectors. The process of datasets, warehouses, or other repositories is known as data mining. Predicting diseases using the vast medical datasets is an extremely difficult task for academics. The researchers employ data mining techniques including classification, clustering, association rules, and others to address this problem. This study&#39;s primary goal is to use classification algorithms to predict liver disorders. Naive Bayes algorithms were employed in this study. Based on their performance characteristics, such as classification accuracy and execution time, these classifier algorithms are contrasted.

Paper 4:
Evaluation of Abnormal Liver Tests [Tinsay A. Woreta 2014]
The diagnosis and treatment of liver illnesses both heavily rely on the use of serum biochemical testing. The routine use of such tests has boosted the diagnosis of liver illnesses in patients who would not otherwise exhibit any symptoms, frequently offering the first indication of liver pathology. In most circumstances, these laboratory tests can assist clinicians in identifying the cause of liver illness in addition to a thorough history, physical examination, and imaging studies. Based on

the degree of aminotransferase increase relative to alkaline phosphatase, liver damage has traditionally been classified as mostly hepatocellular or cholestatic. There is frequently significant overlap in the presentation of different liver disorders, which frequently have a mixed pattern, despite the fact that such a differentiation might help orient early evaluation

Paper 5:

Machine Learning Approaches Binary Classification to Discover Liver Diseases using Clinical Data (*Fahad B. Mostafa* and *Easin Hasan*)

For a medical diagnosis, health professionals use different kinds of pathological ways to make a decision for medical reports in terms of patients’ medical condition. In the modern era, because of the advantage of computers and technologies, one can collect data and visualize many hidden outcomes from them. Statistical machine learning algorithms based on specific problems can assist one to make decisions. Machine learning data driven algorithms can be used to validate existing methods and help researchers to suggest potential new decisions. In this paper, Multiple Imputation by Chained Equations was applied to deal with missing data, and Principal Component Analysis to reduce the dimensionality. To reveal significant findings, data visualizations were implemented. We presented and compared many binary classifier machine learning algorithms (Artificial Neural Network, Random Forest, Support Vector Machine) which were used to classify blood donors and non-blood donors with hepatitis, fibrosis and cirrhosis diseases. From the data published in UCI-MLR, all mentioned techniques were applied to find one better method to classify blood donors and non-blood donors (hepatitis, fibrosis, and cirrhosis) that can help health professionals in a laboratory to make better decisions. Our proposed ML-method showed a better accuracy score (e.g. 98.23% for SVM). Thus, it improved the quality of classification.

Paper 6:

REVIEW OF LIVER DISEASE PREDICTION USING MACHINE LEARNING ALGORITHM (Vijay Panwar, Naved Choudhary, Sonam Mittal)

Liver Disease is the leading cause of global death that impacts the massive quantity of humans around the world. This disease is caused by an assortment of elements that harm the liver. For example, obesity, an undiagnosed hepatitis infection, alcohol

misuse which is responsible for abnormal nerve function, coughing up or vomiting blood, kidney failure, liver failure, jaundice, liver encephalopathy and there are many more. Diagnosis of liver infection at preliminary stage is important for better treatment. In today's scenario devices like sensors are used for detection of infections. Accurate classification techniques are required for automatic identification of disease samples.This disease diagnosis is very costly and complicated. Therefore, the goal of this work is to evaluate the performance of different Machine Learning algorithms in order to reduce the high cost of chronic liver disease diagnosis by prediction. In this work, we used five algorithms Logistic Regression, Decision Tree, Support Vector Machine, Naïve Bayes, and Random Forest. The performance of different classification techniques was evaluated on different measurement techniques such as accuracy, precision, recall, and specificity. We found the accuracy 74%, 72%, 72%, 71%, and 57% for SVM,DT,RF,LR and NB.
The analysis result shown the SVM achieved the highest accuracy. Moreover, our present study mainly focused on the use of clinical data for liver disease prediction and explores different ways of representing such data through our analysis.

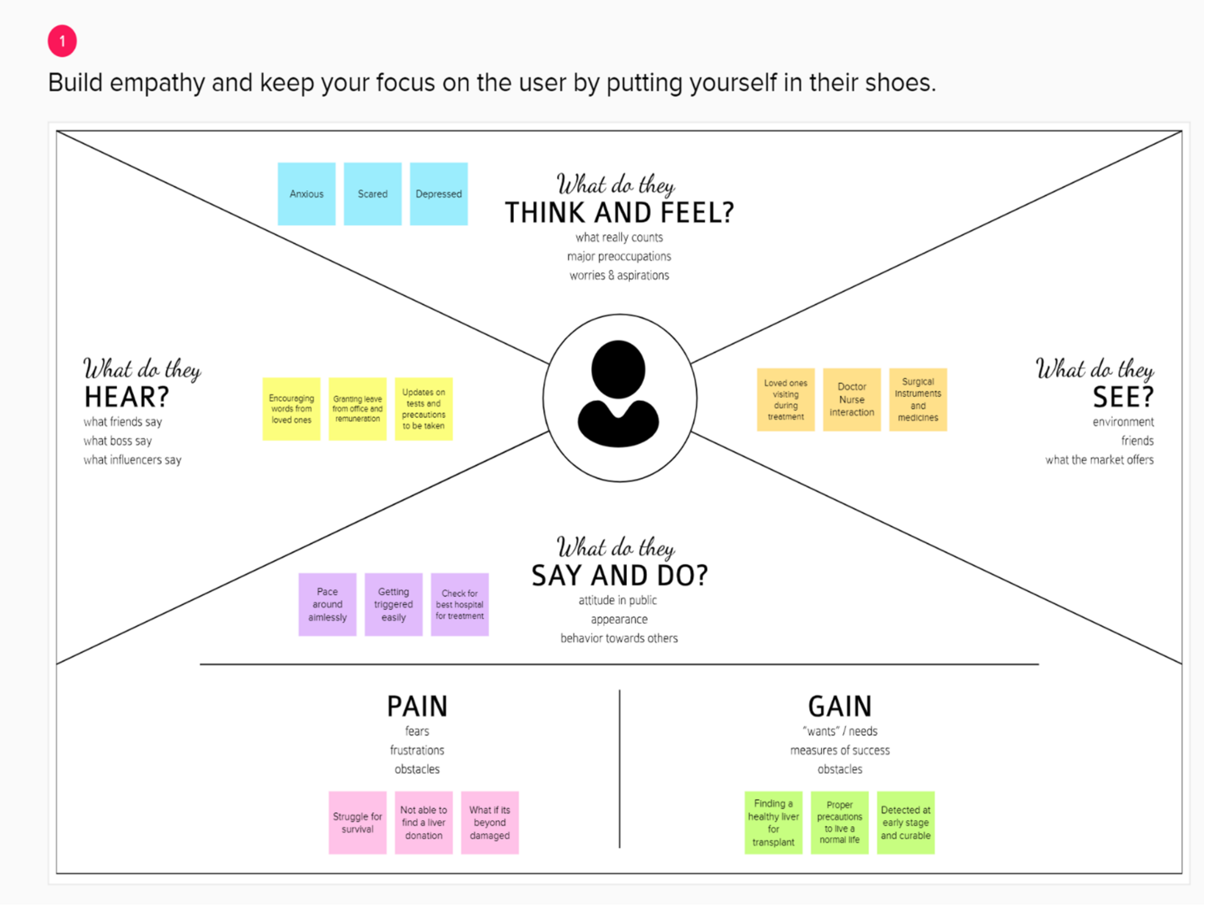
Paper 7:
Performance Analysis of Liver Disease Prediction Using Machine Learning Algorithms (M. Banu Priya, P. Laura Juliet, P.R. Tamilselvi)
Data Mining is one of the most critical aspects of automated disease diagnosis and disease prediction. It involves data mining algorithms and techniques to analyze medical data. In recent years, liver disorders have excessively increased and liver diseases are becoming one of the most fatal diseases in several countries. In this thesis, liver patient datasets are investigate for building classification models in order to predict liver disease. Thisthesis implemented a feature model construction and comparative analysis for improving prediction accuracy of Indian liver patients in three phases. In first phase, min max normalization algorithm is applied on the original liver patient datasets collected from UCI repository. In liver dataset prediction second phase, by the use of PSO feature selection, subset (data) of liver patient dataset from whole normalized liver patient datasets is obtained which comprises only significant attributes. Third phase, classification algorithms are applied on the data set. In the fourth phase, the accuracy will be calculated using root mean Square value, root mean error value.J48 algorithm is considered as the better performance algorithm after applying PSO feature selection. Finally, the evaluation is done based on accuracy values. Thus outputs shows from proposed classification implementations indicate that J48 algorithm performances all other classification algorithm with the help of feature selection with an accuracy of 95.04%.

2.3 PROBLEM STATEMENT DEFINITION

this project aims to identify whether a person has liver disease or not using suitable machine learning algorithm. In order to arrive at the solution, our aim should be to train various machine learning algorithm models on dataset. The data set used is Indian liver patient.csv. so that we have a well performing model which is able to classify any new data as a positive or negative with a reasonable degree of accuracy and perform better

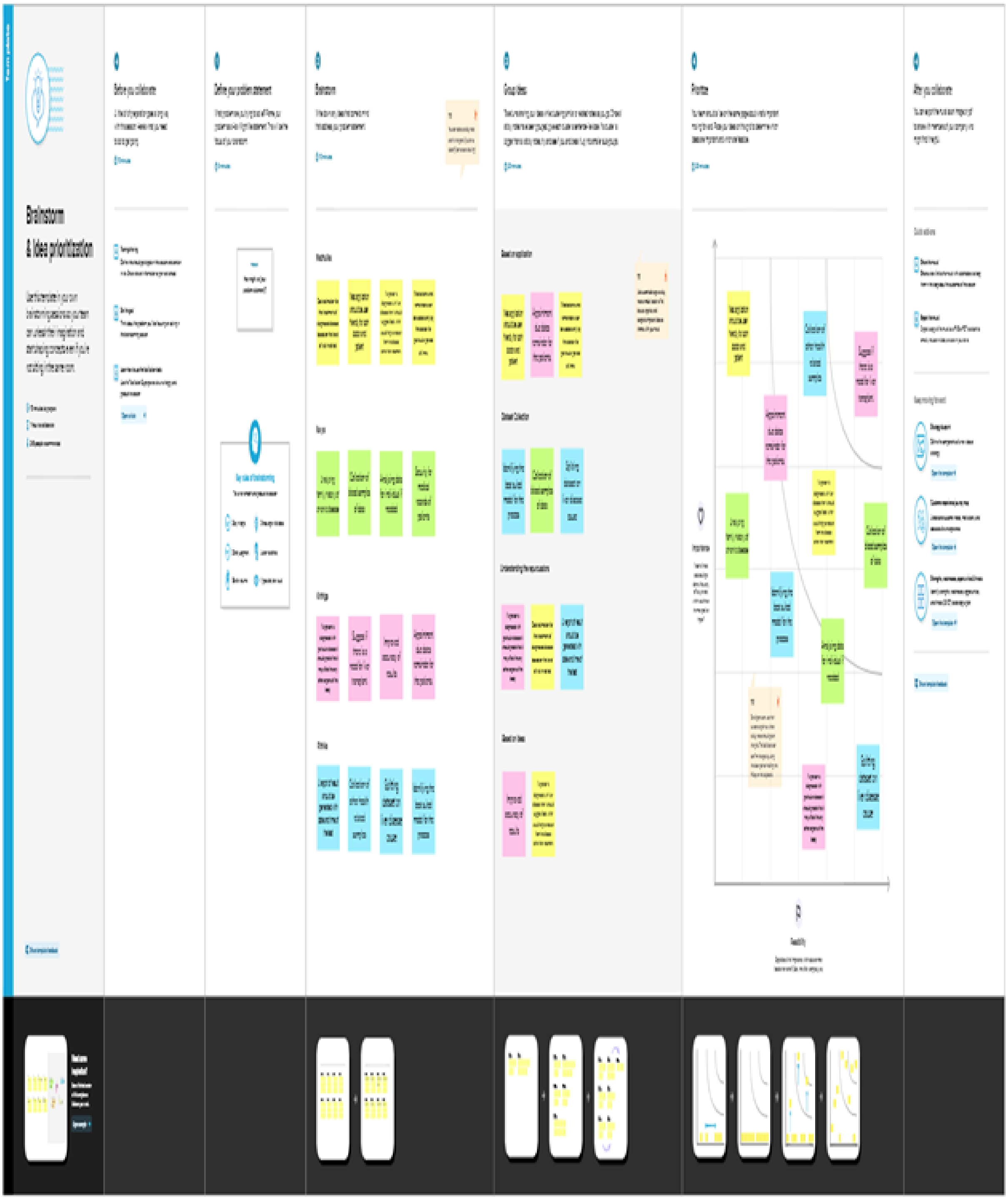
IDEATION &amp; PROPOSED SOLUTION

3.1 EMPATHYMAP CANVAS


1

Build empathy and keep your focus on the user by putting yourself in their shoes.

3.2 IDEATION &amp; BRAINSTORMING



3.3 PROPOSED SOLUTION

S.No.	Parameter	Description
1.	Problem Statement (Problem to be solved)	The challenge is to predict the liver disease patient in faster and accurate way.
2.	Idea / Solution description	We are building a machine learning model which uses statistical data to predict the disease for liver.
3.	Novelty / Uniqueness	The major limitation of CNN is its inability to encode Orientational and relative spatial relationships, view angle. CNN do not encode the position and orientation of data. Lack of ability to be spatially invariant to the input data sample. This is resolved in this research work by combining the genetic algorithm with the CNN method.
4.	Social Impact / Customer Satisfaction	Although knowledge of hepatic biology and pathology is advanced, the prevention and treatment of liver disease lag sadly. This discrepancy is attributable to lack of facilities and

		trained personnel. Morbidity and mortality of liver disease are increasing in frequency because alcoholism, adverse reactions from drug use and abuse, and viral hepatitis are more prevalent. As the nature of these factors suggests, the disadvantaged are particularly at risk.
5.	Business Model (Revenue Model)	It solves the complex process of predicting the liver disease of patients with ease and also provides best results, which in turn helps the doctors to diagnose the liver disease more easily.
6.	Scalability of the Solution	This model can be expanded to include more attributes for more accurate Detection. Can be extended to predict many classification of diseases in early stages.

3.4 PROBLEM SOLUTION FIT

Define CS, fit into CC

1. CUSTOMER SEGMENT(S)

Who is your customer?

i.e. working parents of 0-5 y.o. kids

Patients facing symptoms of liver diseases like abdominal pain and swelling, itchy skin, etc.

Elder people above the age of 60years

6. CUSTOMER CONSTRAINTS

What constraints prevent your customers from taking action or limit their choices

of solutions? i.e. spending power, budget, no cash, network connection, available devices.

Elderly people cannot visit hospitals and medical centers frequently

Patients need to wait for a longer period to get their test reports

5. AVAILABLE SOLUTIONS

or need to get the job done? What have they tried in the past? What pros & cons do these solutions have? i.e. pen and paper is an alternative to digital notetaking

Liver disease diagnosis can be made through any small clinics nearby or through the hospitals

But in both the above

Explore AS, differentiate

2. JOBS-TO-BE-DONE / PROBLEMS

Which jobs-to-be-done (or problems) do you address for your customers? There could be more than one; explore different sides

The solution should diagnose the affected level of liver quicky as possible

9. PROBLEM ROOT CAUSE

What is the real reason that this problem exists? What is the back story behind the need to do this job?

i.e. customers have to do it because of the change in regulations.

- Early diagnosis is beneficial in the treatment of disease

7. BEHAVIOUR

What does your customer do to address the problem

i.e. directly related: find the right solar panel installer, calculate usage and benefits; indirectly associated: customers spend free time on volunteering work (i.e. Greenpeace)

- Patient should consult the doctor if they have symptoms of liver disease.

Focus on J&P, tap into BE, understand RC

3. TRIGGERS

TR

What triggers customers to act? i.e. seeing their neighbour installing solar panels, reading about a more efficient solution in the news.

Understanding the severity of liver disease at later stage by undergoing some severe pains caused as a symptom of liver disease as a symptom of liver disease

Knowing the impact of liver disease through neighbors and friends

4. EMOTIONS: BEFORE / AFTER

EM

How do customers feel when they face a problem or a job and afterwards?

i.e. lost, insecure > confident, in control - use it in your communication strategy & design.

Patients, without knowing that they have been diseased in a particular part of their body might unknowingly do things that are likely to increase the effectiveness of the disease.

10. YOUR SOLUTION

SL

If you are working on an existing business, write down your current solution first, fill in the canvas, and check how much it fits reality.

If you are working on a new business proposition, then keep it blank until you fill in the canvas and come up with a solution that fits within customer limitations, solves a problem and matches customer behaviour.

The solution should give basic recommendation to the patients

The solution should generate the report for the patients for future use

The solution should automatically differentiate healthy and diseased patients just using the data

1. **CHANNELS of BEHAVIOUR**

CH

a. **ONLINE**

What kind of actions do customers take online? Extract online channels from #7

b. **OFFLINE**

What kind of actions do customers take offline? Extract offline channels from #7 and use them for customer development.

Patients need to find the symptoms of liver disease

Patients want to consult the doctor and should follow diagnosis test to predict the liver disease

Patients will be very much cautious about not following some habits after knowing their body condition.		

REQUIREMENT ANALYSIS

4.1 FUNCTIONAL REQUIREMENT

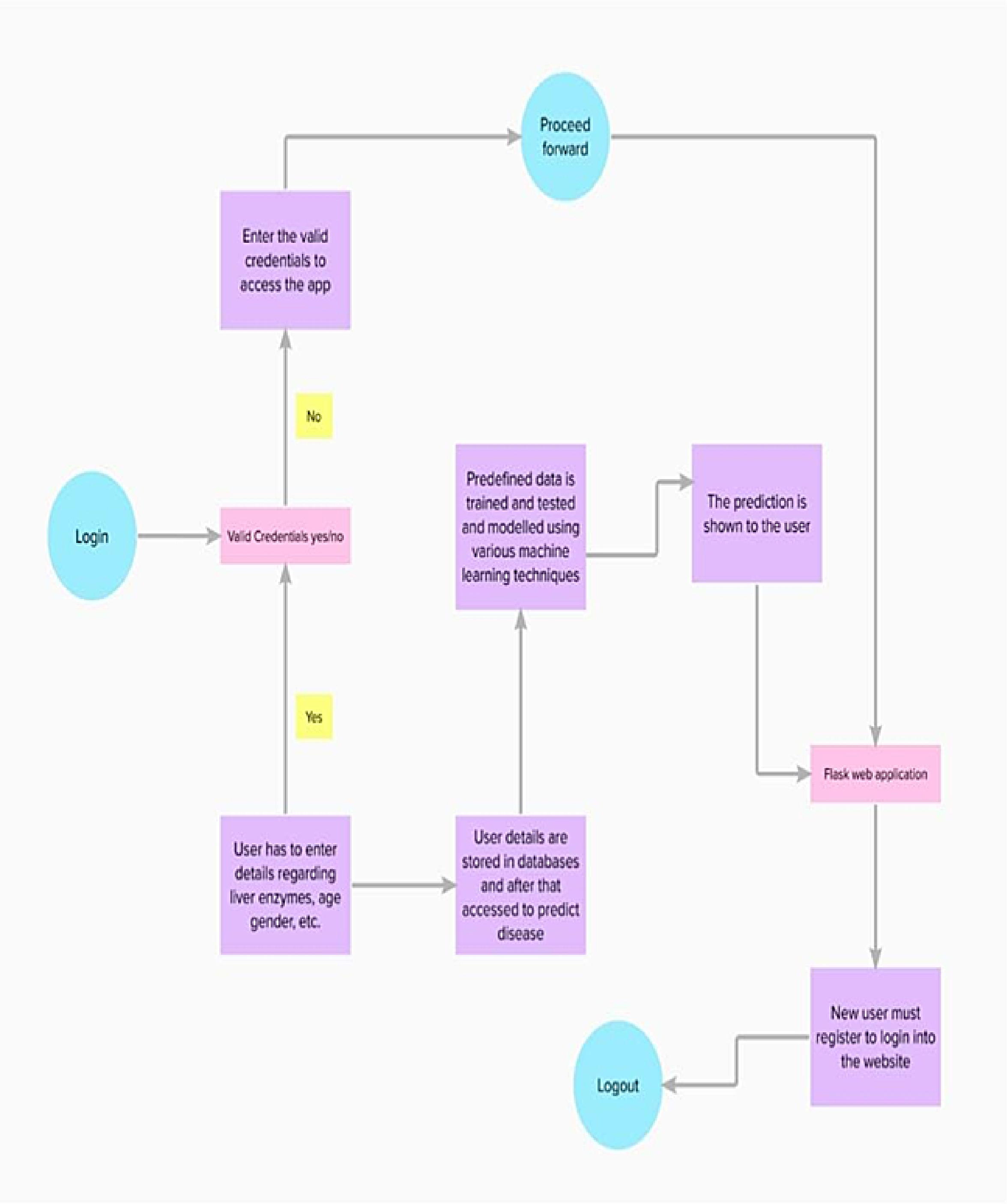
FR No.	Functional Requirement (Epic)	Sub Requirement (Story / Sub- Task)
FR-1	User Registration	Registration through Form present in liver disease prediction website
FR-2	User Confirmation	Confirmation through registered Email
FR-3	Prediction	Based on the data's entered like age, gender and symptoms the type of liver disease is predicted.
FR-4	Hardware Requirements	Intel i3 core processor Internet Connectivity
FR-5	Software Requirements	Windows 7 or higher Python 3.6.0 or higher Visual Studio Code Dataset Jupiter notebook
FR-6	Database Retrieval	we retrieve the data from the database.

4.2 Non-Functional requirements

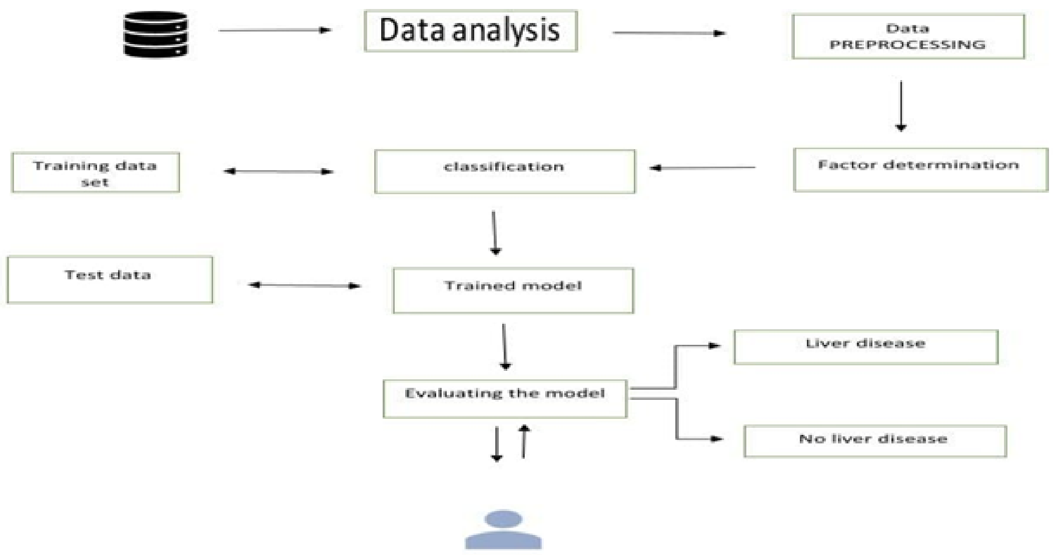
NRF.NO	Non-Functional Requirement	Description
NRF-1	Usability	Due to the early detection of liver disease ,death rate can be decreased
NRF-2	Security	it ensures all data in the system will be Protected
NRF-3	Reliability	it provides secured storing of data and access
NRF-4	Performance	Performance is high as we are using various Machine learning classification algorithms to find the best and the accurate model.
NRF-5	Availability	It can be accessed by all the users.
NRF-6	Scalability	It is acceptable to fit over any place and any resources.

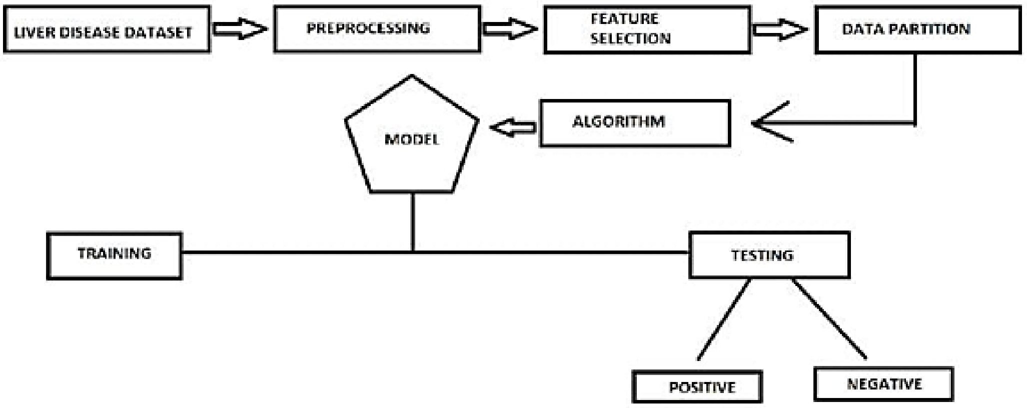
PROJECT DESIGN

5.1 DATA FLOW DIAGRAMS

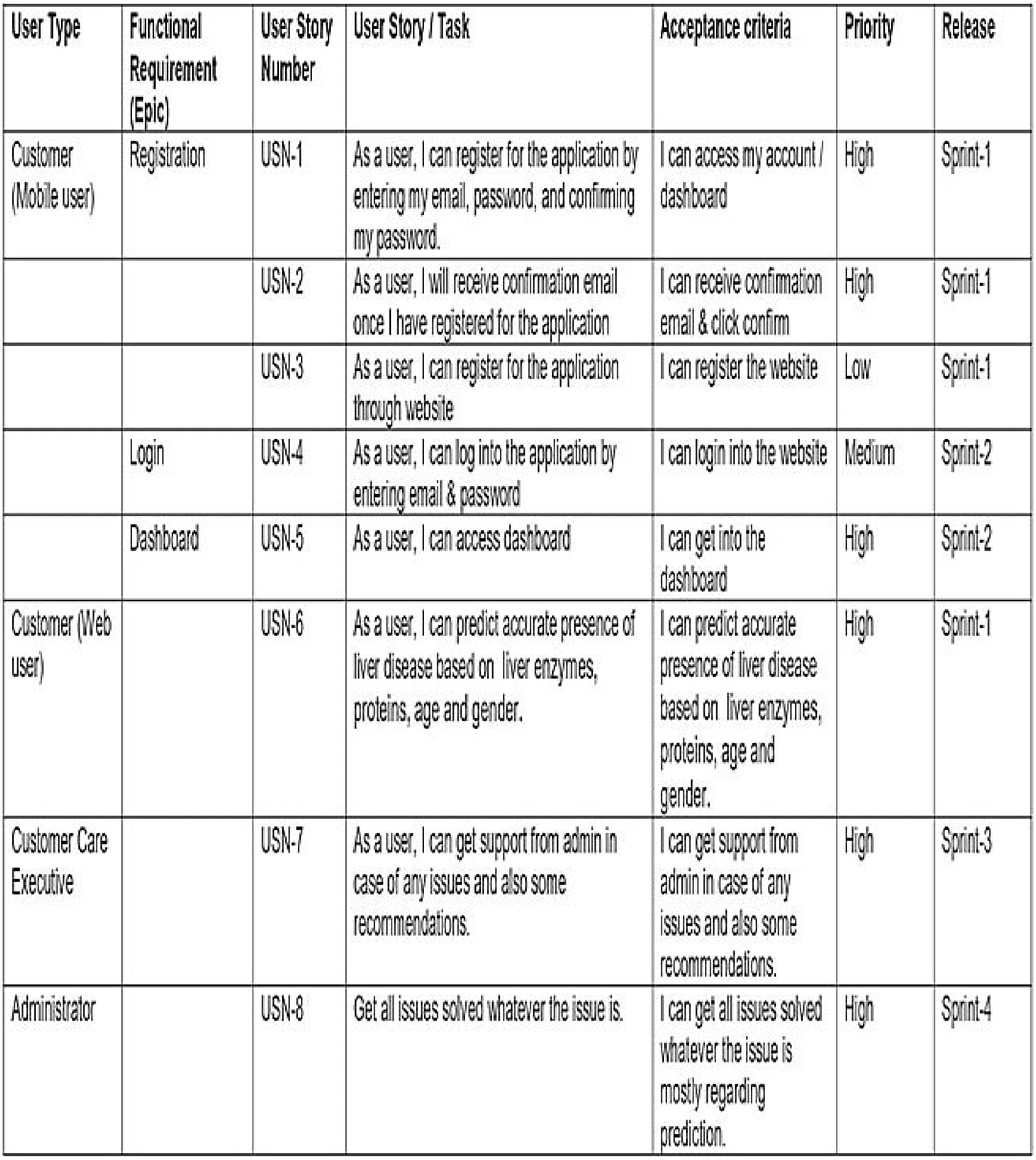

Proceed forward
Enter the valid credentials to access the app
No
Login
Valid Credentials yes/no
Predefined data is trained and tested and modelled using various machine learning techniques
The prediction is shown to the user
Yes
Flask web application
User has to enter details regarding liver enzymes, age gender, etc.
User details are stored in databases and after that accessed to predict disease
New user must register to login into the website
Logout

5.2 SOLUTION ARCHITECTURE





5.3 USER STORIES



User Type	Functional Requirement (Epic)	User Story Number	User Story / Task	Acceptance criteria	Priority	Release
Customer (Mobile user)	Registration	USN-1	As a user, I can register for the application by entering my email, password, and confirming my password.	I can access my account / dashboard	High	Sprint-1
		USN-2	As a user, I will receive confirmation email once I have registered for the application	I can receive confirmation email & click confirm	High	Sprint-1
		USN-3	As a user, I can register for the application through website	I can register the website	Low	Sprint-1
	Login	USN-4	As a user, I can log into the application by entering email & password	I can login into the website	Medium	Sprint-2
	Dashboard	USN-5	As a user, I can access dashboard	I can get into the dashboard	High	Sprint-2
Customer (Web user)		USN-6	As a user, I can predict accurate presence of liver disease based on liver enzymes, proteins, age and gender.	I can predict accurate presence of liver disease based on liver enzymes, proteins, age and gender.	High	Sprint-1
Customer Care Executive		USN-7	As a user, I can get support from admin in case of any issues and also some recommendations.	I can get support from admin in case of any issues and also some recommendations.	High	Sprint-3
Administrator		USN-8	Get all issues solved whatever the issue is.	I can get all issues solved whatever the issue is mostly regarding prediction.	High	Sprint-4

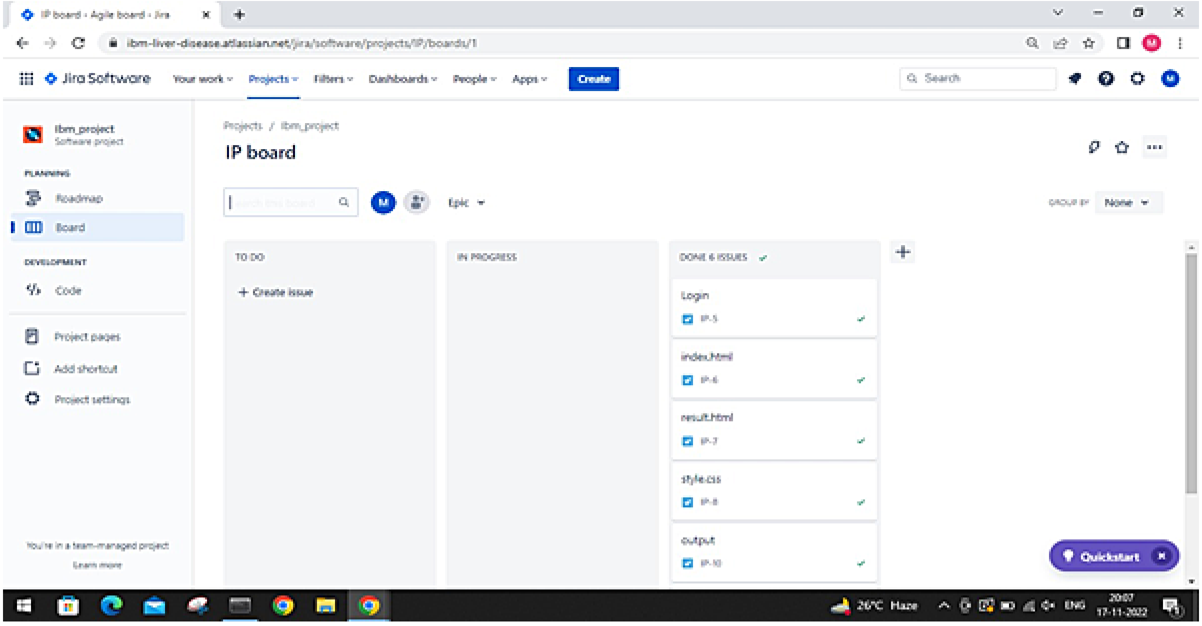
6.1PROJECT PLANNING &amp; SCHEDULING

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	Rithika AM
Sprint-1		USN-2	As a user, I will receive confirmation email once I have registered for the application	5	High	Kavya AP
Sprint-1	login	USN-3	As a user, I can register for the application through email	10	High	Kirthiga S
Sprint-2	Input necessary details	USN-4	As a user, I can give Input Details to Predict.	15	High	Rithika AM
Sprint-2	Pre processing data	USN-5	Transforming the data into suitable format for prediction.	5	High	Madhulika
Sprint -3	Prediction of liver diasease	USN-6	As a user, I can predict Liver Disease using machine learning model.	15	High	Kavya AP
Sprint -3		USN-7	As a user, I can get accurate prediction of liver disease.	10	High	Kirthiga s
Sprint-4	review	USN-8	As a user, I can give feedback of the application.	15	High	Madhulika

a. SPRINT DELIVERY SCHEDULE

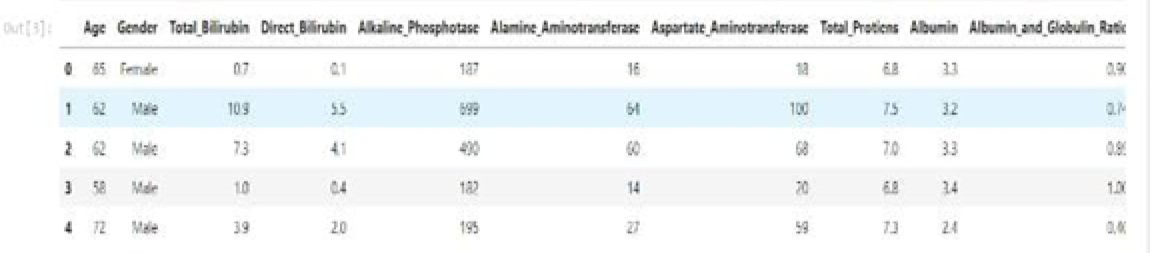
Sprint	Total Story Points	Duration	Sprint Start Date	Sprint End Date (Planned)	Story Points Completed (as on Planned End Date)	Sprint Release Date (Actual)
Sprint-1	20	6 Days	24 Oct 2022	29 Oct 2022	17	29 Oct 2022
Sprint-2	20	6 Days	31 Oct 2022	05 Nov 2022	18	05 Nov 2022
Sprint-3	20	6 Days	07 Nov 2022	12 Nov 2022	17	12 Nov 2022
Sprint-4	20	6 Days	14 Nov 2022	19 Nov 2022	18	19Nov 2022

a. REPORTS FROM JIRA



7.CODING AND SOLUTIONING

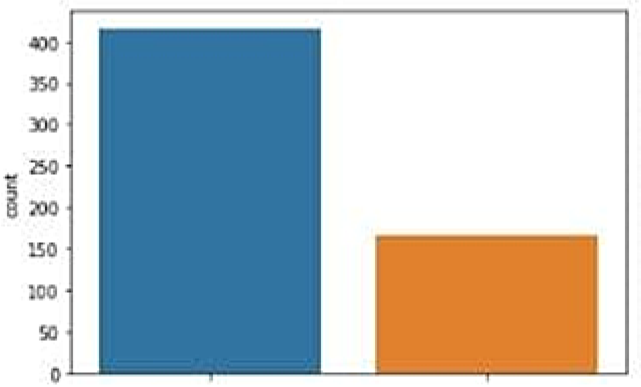
```
data.head()
#to display top five rows of the data
```



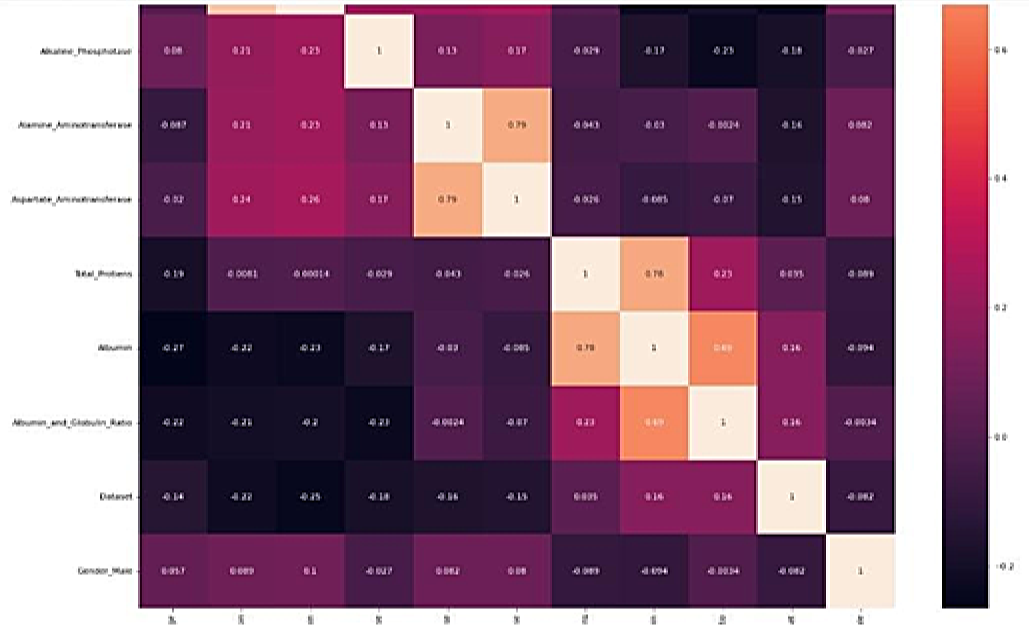
Out[3]:

	Age	Gender	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphotase	Alamine_Aminotransferase	Aspartate_Aminotransferase	Total_Protiens	Albumin	Albumin_and_Globulin_Ratic
0	65	Female	0.7	0.1	187	16	18	6.8	3.3	0.9
1	62	Male	10.9	5.5	699	64	100	7.5	3.2	0.7
2	62	Male	7.3	4.1	490	60	68	7.0	3.3	0.8
3	58	Male	1.0	0.4	182	14	20	6.8	3.4	1.0
4	72	Male	3.9	2.0	195	27	59	7.3	2.4	0.6

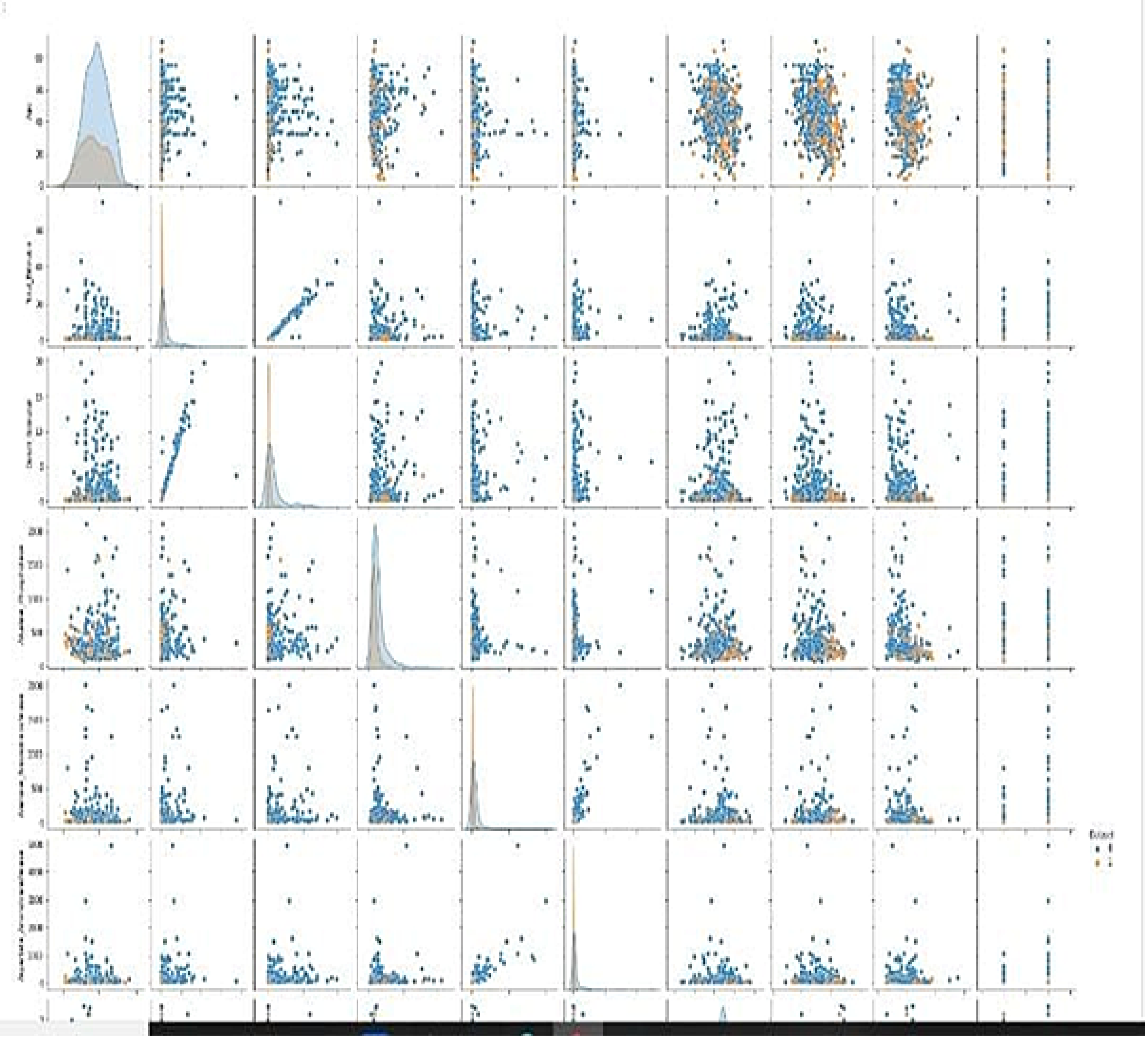
```
import seaborn as sns
sns.countplot('Dataset', data = data)
```



```
plt.figure(figsize = (20,20))
sns.heatmap(data.corr(), annot =True)
```



```
sns.pairplot(data, hue = 'Dataset')
```



8.TEST CASE

Testing forms an integral part of any software development project. Testing helps in ensuring that the final product is by and large, free of defects and it meets the desired requirements. Proper testing in the development phase helps in identifying the critical errors in the design and implementation of various functionalities thereby ensuring product reliability

1)Pre-train tests

The intention is to write such tests which can be run without trained parameters so that we can catch implementation errors early on. This helps in avoiding the extra time and effort spent in a wasted training job

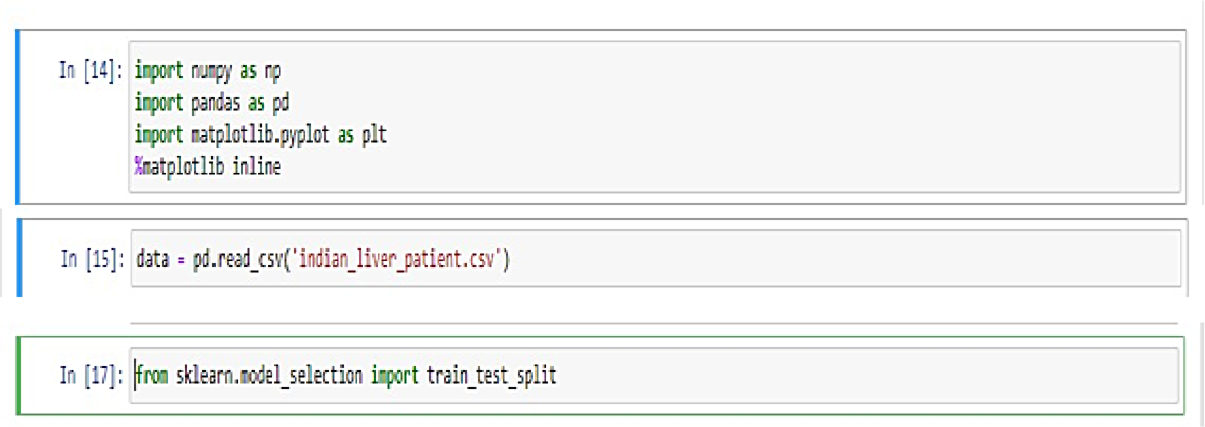
We can test the following in the pre-train test:

- test dataset leakage i.e. checking whether the data in training and testing datasets have no duplication
- check for the output ranges. In the cases where we are predicting outputs in a certain range (for example when predicting probabilities), we need to ensure the final prediction is not outside the expected range of values.

Post-train tests: Post-train tests are aimed at testing the model's behavior. We want to test the learned logic and it could be tested on the following points and more:

- invariance tests which involve testing the model by tweaking only one feature in a data point and checking for consistency in model predictions..
- Directional expectations wherein we test for a direct relation between feature values and predictions

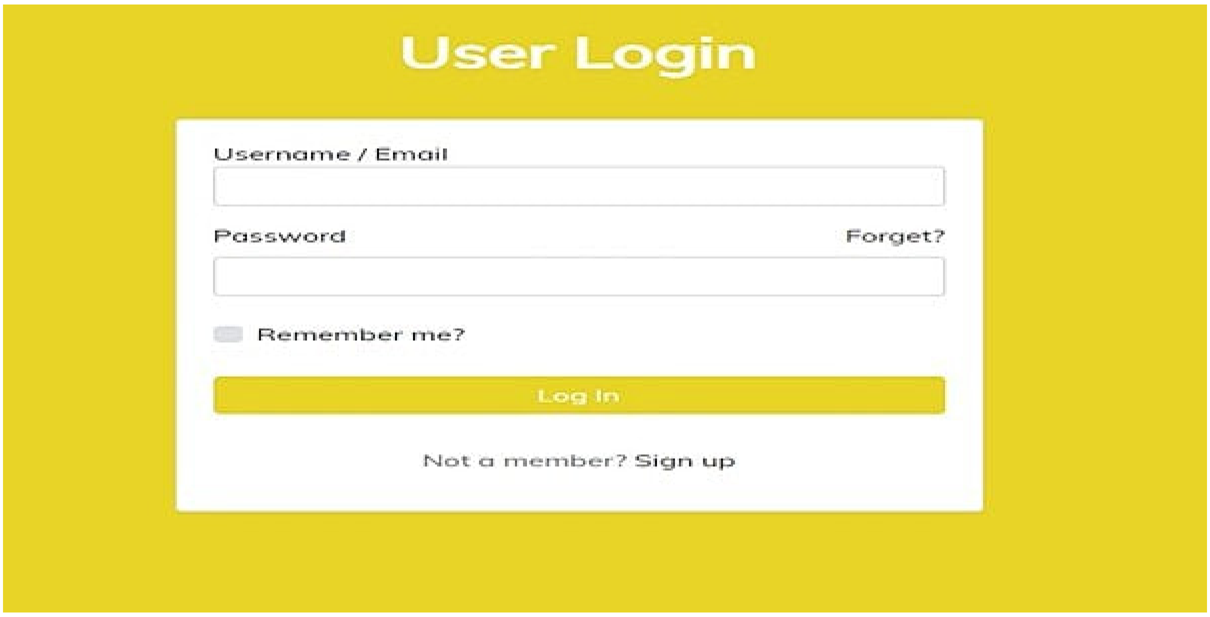
Helper Functions Functions for loading data:



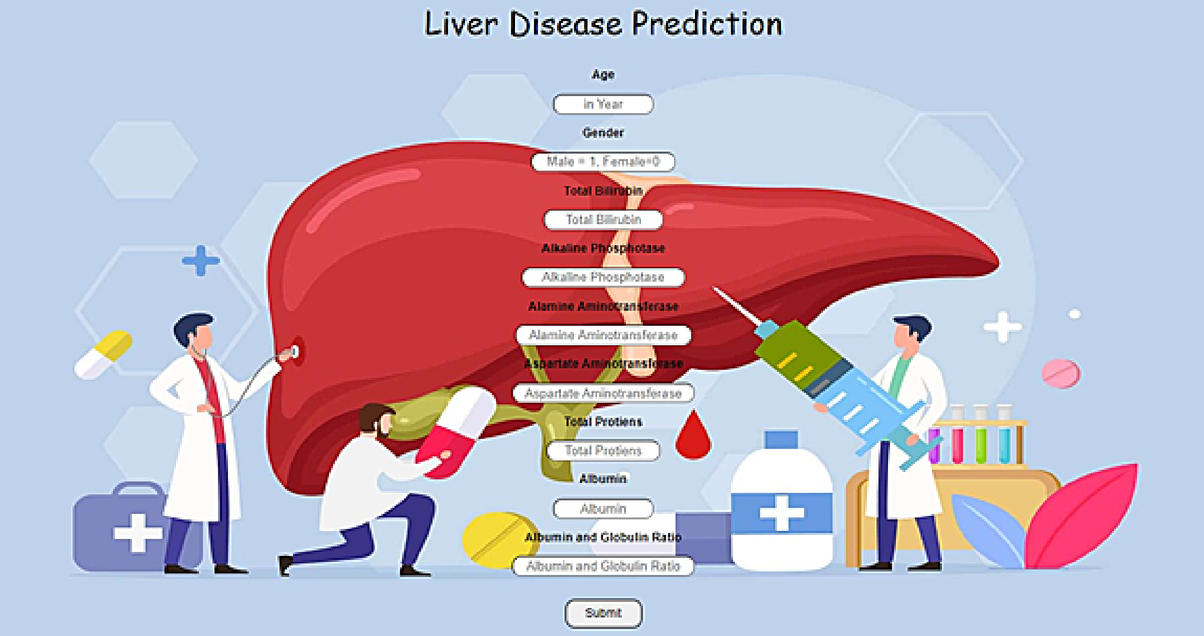
```
In [14]: import numpy as np
         import pandas as pd
         import matplotlib.pyplot as plt
         %matplotlib inline

In [15]: data = pd.read_csv('indian_liver_patient.csv')

In [17]: from sklearn.model_selection import train_test_split
```

RESULTS

Website login page:



Liver disease prediction :



Liver Disease Prediction

Age

in Year

Gender

Male = 1, Female=0

Total Billirubin

Total Billirubin

Alkaline Phosphotase

Alkaline Phosphotase

Alamine Aminotransferase

Alamine Aminotransferase

Aspartate Aminotransferase

Aspartate Aminotransferase

Total Protiens

Total Protiens

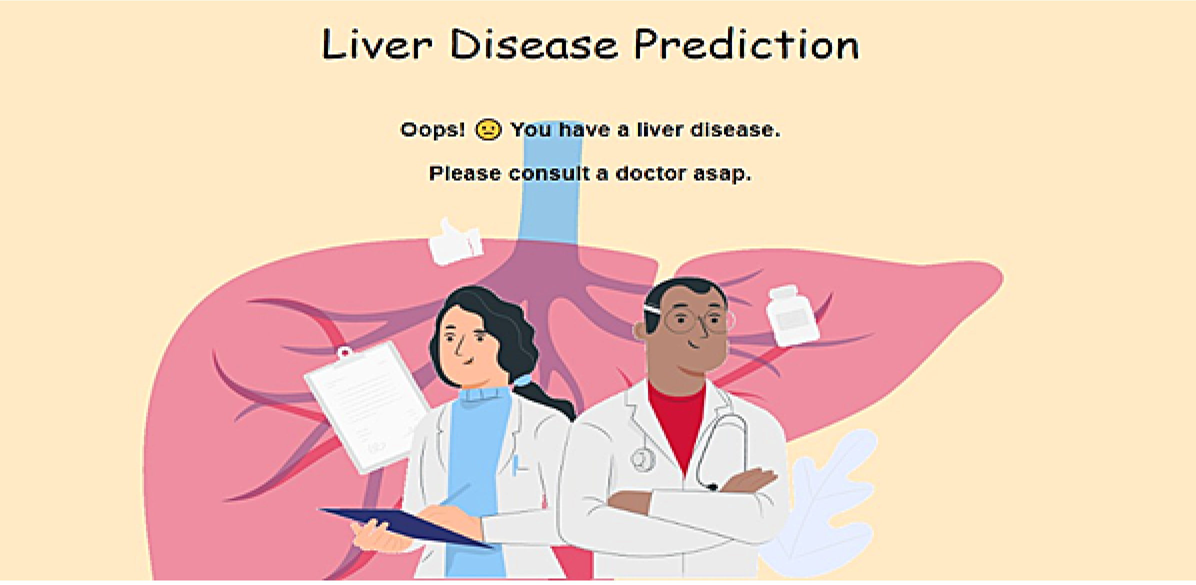
Albumin

Albumin

Albumin and Globulin Ratio

Albumin and Globulin Ratio

Liver disease predicted

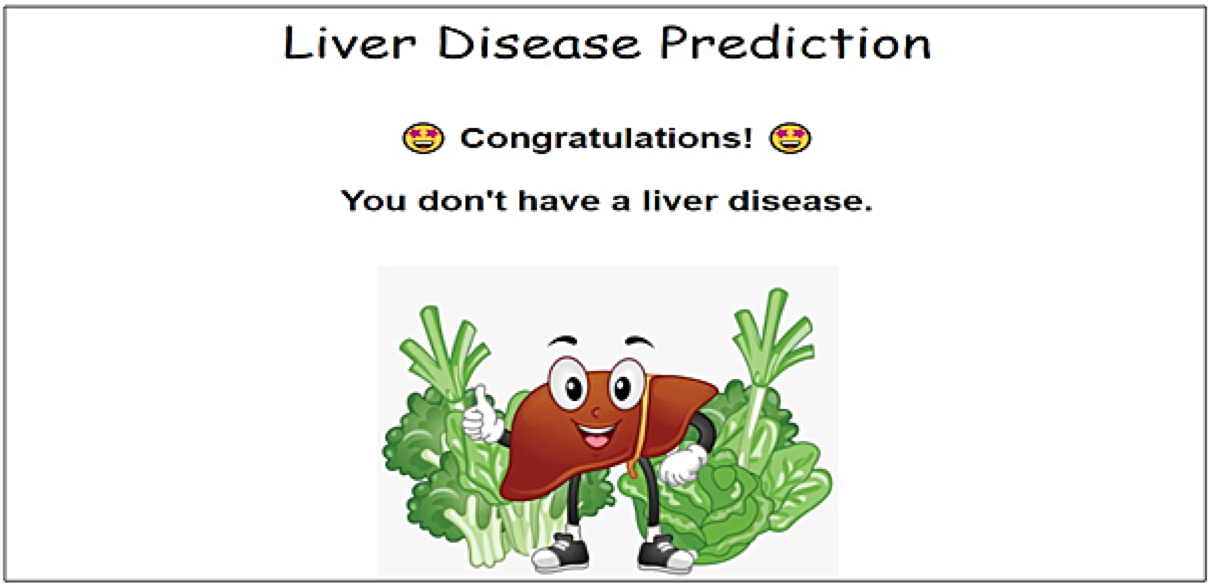


Liver Disease Prediction

Oops! 😕 You have a liver disease.

Please consult a doctor asap.

No liver disease predicted



Liver Disease Prediction

🤩 Congratulations! 🤩

You don't have a liver disease.

ADVANTAGES & DISADVANTAGES

Advantage:

1.No medical expertise required: You dont need to have any knowledge of medical
science and liver diseases to predict the liver disease using this application. All you need to do is enter the details being asked, which are already present in the blood test report(some like age, gender are already known) and then you will get the results of prediction.

2.Immediate results: The results here are predicted within seconds of entering the details. You dont need to wait for a doctor to come, unlike in traditional method.

Disadvantage:

There may occur wrong prediction of liver disease due to data try by patients.

CONCLUSION:

Diseases related to liver and heart are becoming more and more common with time. With continuous technological advancements, these are only going to increase in the future. Although people are becoming more conscious of health nowadays and are joining yoga classes, dance classes; still the sedentary lifestyle and luxuries that are continuously being introduced and enhanced; the problem is going to last long.

So, in such a scenario, our project will be extremely helpful to the society. With the dataset that we used for this project, we got 76%accuracy, and though it might be difficult to get such accuracies with very large datasets, from this projects results, one can clearly conclude that we can predict the risk of liver diseases with accuracy of 90 % or more.

12. FUTURE SCOPE

In this project the proposed system we have choosen Indian Liver Patient Dataset. We analysed the liver disease using algorithms such as Random Forest, and Bayesnet Classification.

There are many criterions for evaluating the selected feature subset, here this thesis used features such as Total bilirubin, Direct_ bilirubin, Alkaline_Phosphotase, Alamine_Aminotransferase, Aspartate_Aminotransferase, Total_Protiens, Albumin to evaluate the performance of different classification algorithm. In future, we have attempted to classify different feature selection algorithms into four groups: complete search, heuristic search, meta-heuristic methods and methods that use artificial neural network.

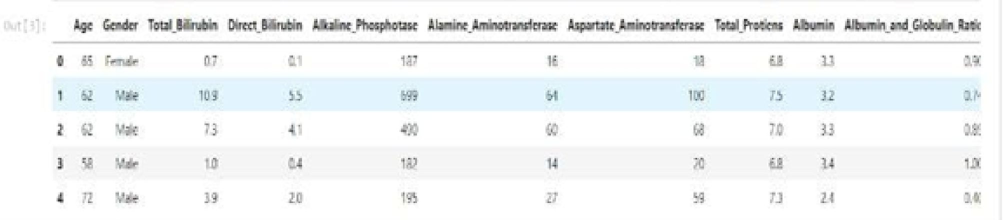
There is a scope to further reduce search space for better liver classification accuracy if enhanced selection and mutation procedures are being used. The future methodology is used to analyse the liver region into separable compartments i.e. liver etc. However, the method requires further improvement mostly regarding feature selection of the liver into multiple components: renal cortex, renal column, renal medulla and renal pelvis. Apart from that, it is planned to expand the database on which the system will be tested. And also the proposed method in this project can be employed for detecting the heart diseases in future with the heart dataset and classification of the diseases.

This Disease Prediction system can be used for urgent guidance on their illness according to the details and symptoms they will feed to the web-based application. Here, some intelligent
data processing techniques are used to get the most accurate disease that would be related to the patient's details

APPENDIX

Source code:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
#Importing libraries
data = pd.read_csv('indian_liver_patient.csv')
#reading CSV file using pandas
data.head()
#to display top five rows of the data
```

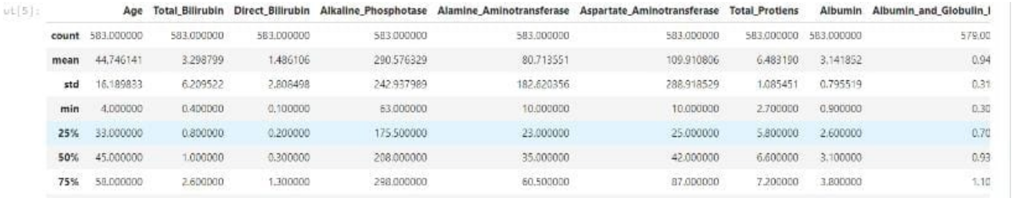


Out[3]:

	Age	Gender	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphotase	Alamine_Aminotransferase	Aspartate_Aminotransferase	Total_Protiens	Albumin	Albumin_and_Globulin_Ratic
0	65	Female	0.7	0.1	187	16	18	6.8	3.3	0.9
1	62	Male	10.9	5.5	699	64	100	7.5	3.2	0.7
2	62	Male	7.3	4.1	490	60	68	7.0	3.3	0.8
3	58	Male	1.0	0.4	182	14	20	6.8	3.4	1.0
4	72	Male	3.9	2.0	195	27	59	7.3	2.4	0.4

```
data.info()

RangeIndex: 583 entries, 0 to 582
Data columns (total 11 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   Age                          583 non-null    int64
 1   Gender                       583 non-null    object
 2   Total_Bilirubin              583 non-null    float64
 3   Direct_Bilirubin             583 non-null    float64
 4   Alkaline_Phosphotase         583 non-null    int64
 5   Alamine_Aminotransferase     583 non-null    int64
 6   Aspartate_Aminotransferase   583 non-null    int64
 7   Total_Protiens               583 non-null    float64
 8   Albumin                      583 non-null    float64
 9   Albumin_and_Globulin_Ratio   579 non-null    float64
 10  Dataset                      583 non-null    int64
dtypes: float64(5), int64(5), object(1)
memory usage: 50.2+ K

data.describe()
```

#The data types in pandas dataframes are the object, float, int64, bool, and datetime64. We should know the data type of each column.

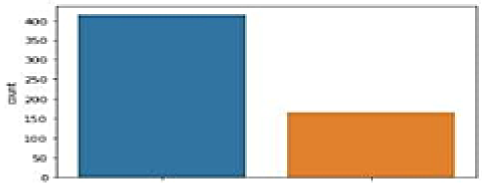


ut[5]:

	Age	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphotase	Alamine_Aminotransferase	Aspartate_Aminotransferase	Total_Protiens	Albumin	Albumin_and_Globulin_I
count	583.000000	583.000000	583.000000	583.000000	583.000000	583.000000	583.000000	583.000000	579.00
mean	44.746141	3.298799	1.486106	290.576329	80.713551	109.910806	6.483190	3.141852	0.94
std	16.189833	6.209522	2.808498	242.937989	182.620356	288.918529	1.085451	0.795519	0.31
min	4.000000	0.400000	0.100000	63.000000	10.000000	10.000000	2.700000	0.900000	0.30
25%	33.000000	0.800000	0.200000	175.500000	23.000000	25.000000	5.800000	2.600000	0.70
50%	45.000000	1.000000	0.300000	208.000000	35.000000	42.000000	6.600000	3.100000	0.93
75%	58.000000	2.600000	1.300000	298.000000	60.500000	87.000000	7.200000	3.800000	1.10

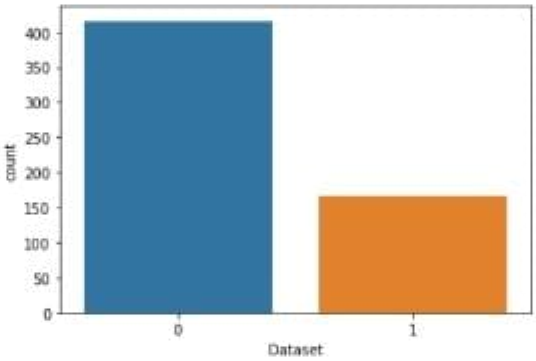
```
data.isnull().sum()
```

```
                  ata.isnull().sum()

Age                            0
Gender                         0
Total_Bilirubin                0
Direct_Bilirubin               0
Alkaline_Phosphotase           0
Alamine_Aminotransferase       0
Aspartate_Aminotransferase     0
Total_Protiens                 0
Albumin                        0
Albumin_and_Globulin_Ratio     4
Dataset                        0
dtype: int64
```

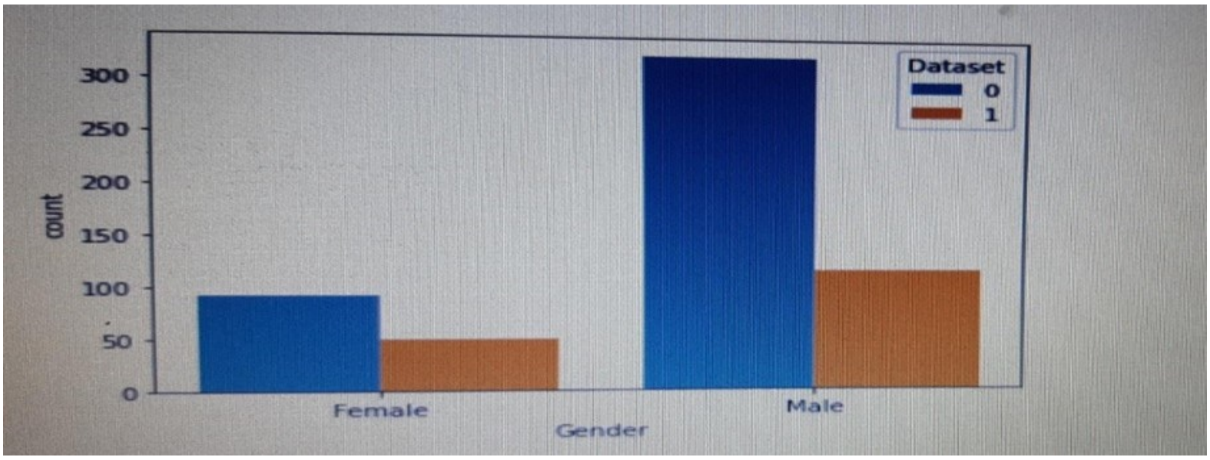


```
 data['Dataset'] = data['Dataset'].replace([2,1],[1,0])
data['Dataset'].head()
0     0
1     0
2     0
3     0
4     0
Name: Dataset, dtype: int64
import seaborn as sns
sns.countplot('Dataset', data = data)
```

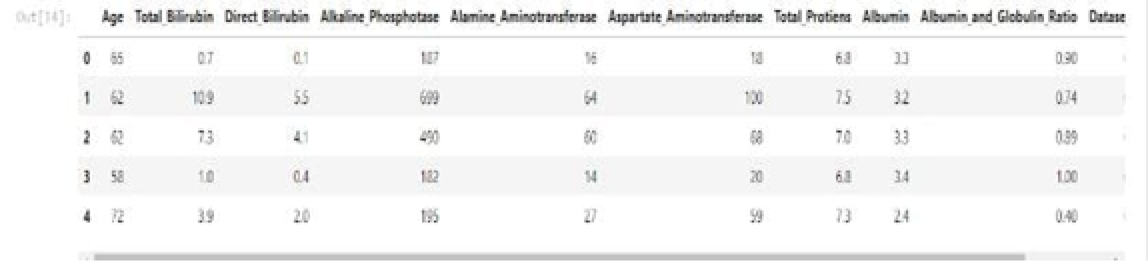
Out[11]:



```
sns.countplot('Gender', data = data, hue = 'Dataset')
```



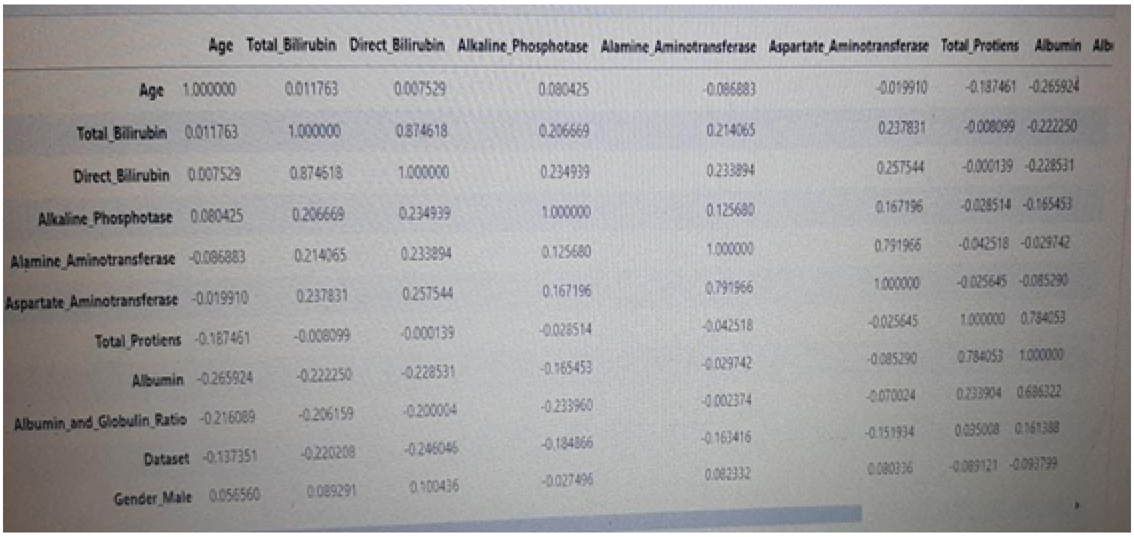
```
sns.countplot('Gender', data = data, hue = 'Dataset')
```



Out[14]:

	Age	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphotase	Alamine_Aminotransferase	Aspartate_Aminotransferase	Total_Protiens	Albumin	Albumin_and_Globulin_Ratio	Datase
0	65	0.7	0.1	187	16	18	6.8	3.3	0.90	
1	62	10.9	5.5	699	64	100	7.5	3.2	0.74	
2	62	7.3	4.1	490	60	68	7.0	3.3	0.89	
3	58	1.0	0.4	182	14	20	6.8	3.4	1.00	
4	72	3.9	2.0	195	27	59	7.3	2.4	0.40	

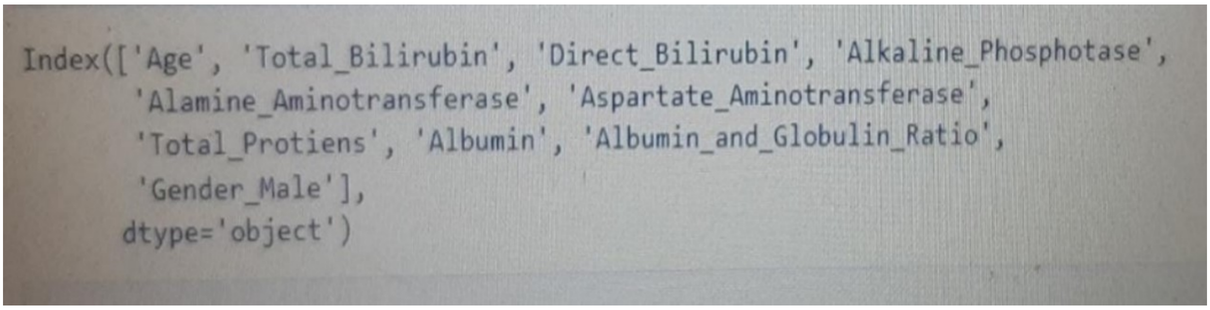
```
plt.figure(figsize = (20,20))
sns.heatmap(data.corr(), annot =True)
sns.pairplot(data, hue = 'Dataset')
data.corr()
```



	Age	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphotase	Alamine_Aminotransferase	Aspartate_Aminotransferase	Total_Protiens	Albumin	Alb
Age	1.000000	0.011763	0.007529	0.080425	-0.086883	-0.019910	-0.187461	-0.265924	
Total_Bilirubin	0.011763	1.000000	0.874618	0.206669	0.214065	0.237831	-0.008099	-0.222250	
Direct_Bilirubin	0.007529	0.874618	1.000000	0.234939	0.233894	0.257544	-0.000139	-0.228531	
Alkaline_Phosphotase	0.080425	0.206669	0.234939	1.000000	0.125680	0.167196	-0.028514	-0.165453	
Alamine_Aminotransferase	-0.086883	0.214065	0.233894	0.125680	1.000000	0.791966	-0.042518	-0.029742	
Aspartate_Aminotransferase	-0.019910	0.237831	0.257544	0.167196	0.791966	1.000000	-0.025645	-0.085290	
Total_Protiens	-0.187461	-0.008099	-0.000139	-0.028514	-0.042518	-0.025645	1.000000	0.784053	
Albumin	-0.265924	-0.222250	-0.228531	-0.165453	-0.029742	-0.085290	0.784053	1.000000	
Albumin_and_Globulin_Ratio	-0.216089	-0.206159	-0.200004	-0.233960	-0.002374	-0.070024	0.233904	0.686322	
Dataset	-0.137351	-0.220208	-0.246046	-0.184866	-0.163416	-0.151934	0.035008	0.161388	
Gender_Male	0.056560	0.089291	0.100436	-0.027496	0.082332	0.080336	-0.089121	-0.093799	

```
# X = data[['Albumin_and_Globulin_Ratio', 'Albumin', 'Total_Protiens',
'Aspartate_Aminotransferase', 'Alamine_Aminotransferase',
'Alkaline_Phosphotase', 'Age']]
X = data.drop('Dataset', axis = 1)
y = data['Dataset']
```

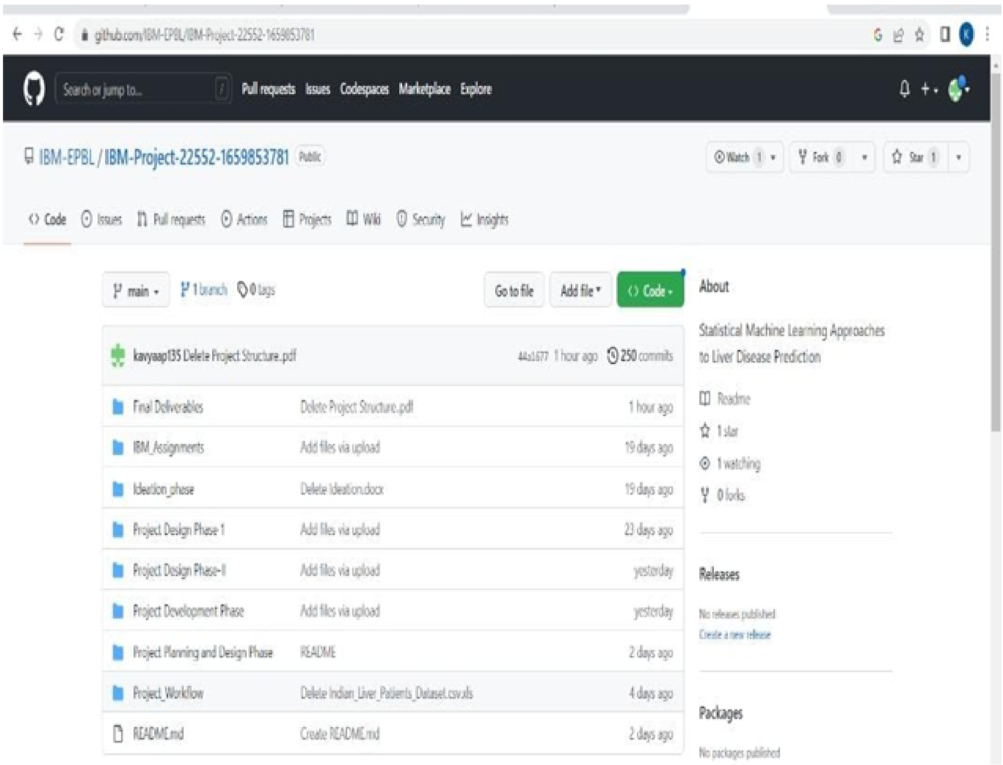
```
X.columns
```



```
Index(['Age', 'Total_Bilirubin', 'Direct_Bilirubin', 'Alkaline_Phosphotase',
       'Alamine_Aminotransferase', 'Aspartate_Aminotransferase',
       'Total_Protiens', 'Albumin', 'Albumin_and_Globulin_Ratio',
       'Gender_Male'],
      dtype='object')
```

```
from sklearn.model _selection import train_test_split
  X_train, X_test, y _train, y_test = train _test _split (X, y, test _size
= 0.1,       random_state = 42)
print("Train Set: ", X_train.shape, y_train.shape)
print("Test Set: ", X_test.shape, y_test.shape)

Train Set:  (524, 10) (524,)
Test Set:  (59, 10) (59,)
from sklearn.ensemble import RandomForestClassifier
model = RandomForestClassifier(n_estimators=20)
model.fit(X_train, y_train)
RandomForestClassifier(n_estimators=20)
from sklearn.metrics import confusion_matrix, accuracy_score
confusion_matrix(y_test, model.predict(X_test))
array([[40,  5],
       [ 8,  6]], dtype=int64)
print(f"Accuracy is {round(accuracy_score(y_test,
model.predict(X_test))*100,2)}")
Accuracy is 77.97
Import pickle
Pickle.dump(model,open("liver.pkl",'wb'))
```

Github link:
https://github.com/IBM-EPBL/IBM-Project-22552-1659853781



kavyaap135 Delete Project Structure.pdf 44a1677 1 hour ago 250 commits

Final Deliverables	Delete Project Structure.pdf	1 hour ago
IBM_Assignments	Add files via upload	19 days ago
Ideation_phase	Delete Ideation.docx	19 days ago
Project Design Phase 1	Add files via upload	23 days ago
Project Design Phase-II	Add files via upload	yesterday
Project Development Phase	Add files via upload	yesterday
Project Planning and Design Phase	README	2 days ago
Project_Workflow	Delete Indian_Liver_Patients_Dataset.csv.xls	4 days ago
README.md	Create README.md	2 days ago

About

Statistical Machine Learning Approaches to Liver Disease Prediction

Releases

No releases published

Packages

No packages published

```
from sklearn.model _selection import train_test_split
  X_train, X_test, y _train, y_test = train _test _split (X, y, test _size
= 0.1,       random_state = 42)
print("Train Set: ", X_train.shape, y_train.shape)
print("Test Set: ", X_test.shape, y_test.shape)

Train Set:  (524, 10) (524,)
Test Set:  (59, 10) (59,)
```