PROJECT DOCUMENTATION

1. INTRODUCTION

1.1 Project Overview

1.2 Purpose

2. LITERATURE SURVEY

2.1 Existing problem

2.2 References

2.3 Problem Statement Definition

3. IDEATION & PROPOSED SOLUTION

3.1 Empathy Map Canvas

3.2 Ideation & Brainstorming

3.3 Proposed Solution

3.4 Problem Solution fit

4. REQUIREMENT ANALYSIS

4.1 Functional requirement

4.2 Non-Functional requirements

5. PROJECT DESIGN

5.1 Data Flow Diagrams

5.2 Solution & Technical Architecture

5.3 User Stories

6. PROJECT PLANNING & SCHEDULING

6.1 Sprint Planning & Estimation

6.2 Sprint Delivery Schedule

6.3 Reports from JIRA

7. CODING & SOLUTIONING (Explain the features added in the project along with code)

7.1 Feature 1

7.2 Feature 2

7.3 Database Schema (if Applicable)

8. TESTING

8.1 Test Cases

8.2 User Acceptance Testing

9. RESULTS

9.1 Performance Metrics

10. ADVANTAGES & DISADVANTAGES

11. CONCLUSION

12. FUTURE SCOPE

13. APPENDIX

Source Code

GitHub & Project Demo Link

**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. **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'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'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.

Paper 2 :

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'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'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'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 illnessesin 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

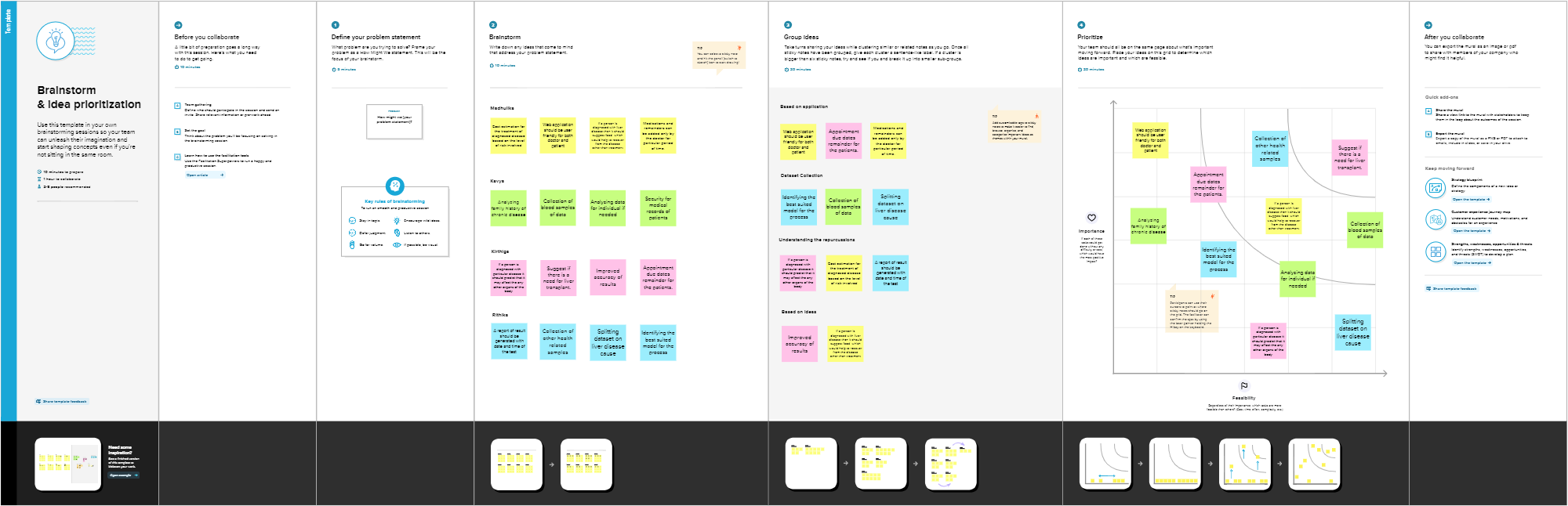
this project aims to identify whether a person has liver disease or not using suitable machine learning algorithm. In order to arrive at the solution, our aim should be to train various machine learning algorithm models on dataset. The data set used is Indian liver patient.csv. so that we have a well 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 & PROPOSED SOLUTION**

3.1 EMPATHYMAP CANVAS



**3.2 IDEATION & BRAINSTORMING**



**3.3 PROPOSED SOLUTION**

|  |  |  |
| --- | --- | --- |
| **S.No.** | **Parameter** | **Description** |
|  | Problem Statement (Problem to be solved) | The challenge is to predict the liver disease patient in faster and accurate way. |
|  | Idea / Solution description | We are building a machine learning model which uses statistical data to predict the disease for liver. |
|  | 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. |
|  | 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. |
|  | 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. |
|  | 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

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 cases, patient will wait for a longer period to get their test results during which the disease may get severe

**AS**

**5. AVAILABLE SOLUTIONS**

Which solutions are available to the customers when they face the problem

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
* Liver disease diagnosis test is not affordable for poor people

**CC**

**6. CUSTOMER CONSTRAINTS**

**CS**

**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
* Hospitals
* Liver Disease diagnosis centers.

**Explore AS, differentiate**

**Define CS, fit into CC**

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.
* Patient want to follow liver disease diagnosis test on a monthly basis as per doctor’s consultancy.

**BE**

**7. BEHAVIOUR**

What does your customer do to address the problem and get the job done?

**RC**

**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
* People mostly hesitate to visit hospitals or they may not be knowing the severity of the disease

**J&P**

**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
* Solution should generate a detail medical report

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

**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  Patients will be very much cautious about not following some habits after knowing their body condition | **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**     1. **ONLINE**   What kind of actions do customers take online? Extract online channels from #7   * 1. **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 |

**REQUIREMENT ANALYSIS**

4.1 FUNCTIONAL REQUIREMENT

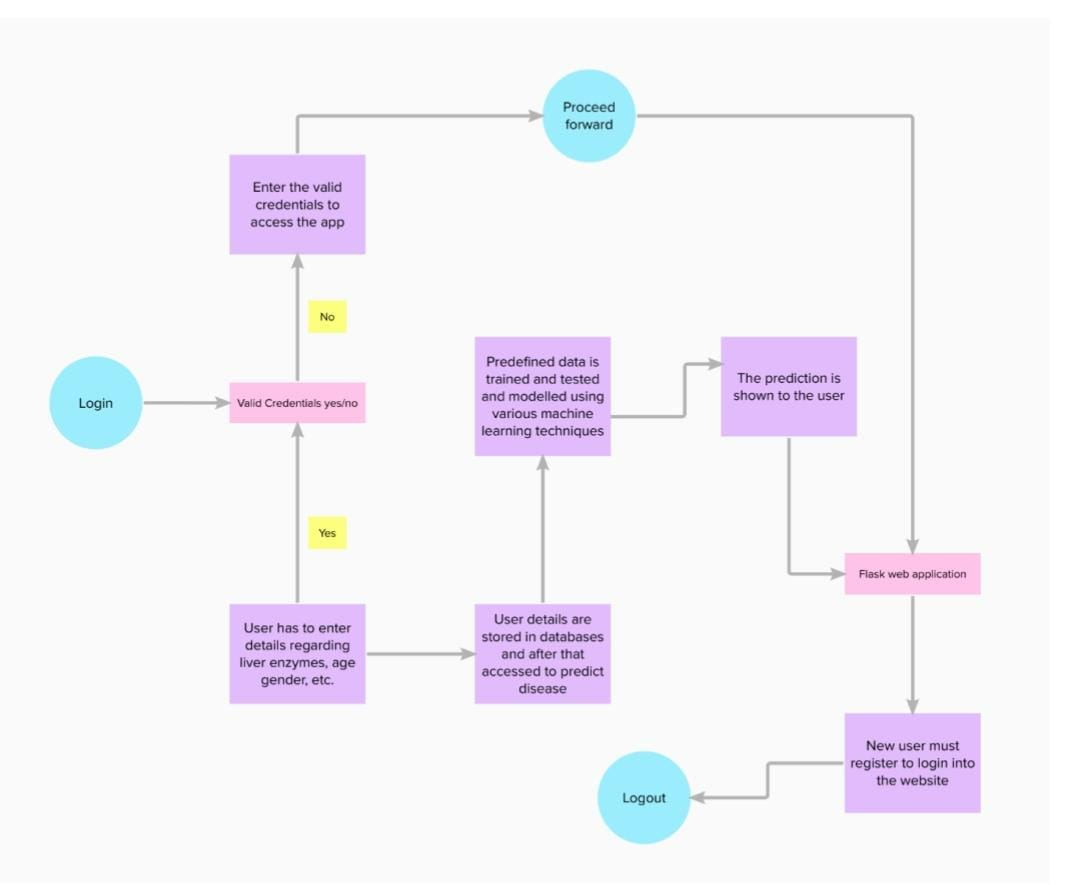
|  |  |  |
| --- | --- | --- |
| FR No. | Functional Requirement (Epic) | Sub Requirement (Story / Sub-Task) |
| **FR-1** | User Registration | Registration through Form present in liver disease prediction website |
| **FR-2** | User Confirmation | Confirmation through registered Email |
| **FR-3** | 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

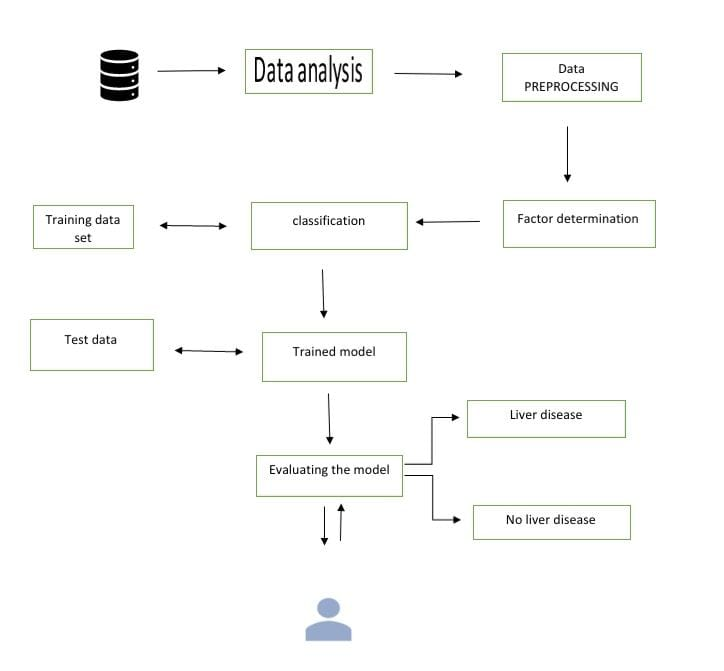
|  |  |  |
| --- | --- | --- |
| **NRF.NO** | Non-Functional Requirement | Description |
| **NRF-1** | Usability | Due to the early detection of liver disease ,death rate can be decreased |
| **NRF-2** | Security | it ensures all data in the system will be Protected |
| **NRF-3** | Reliability | it provides 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**

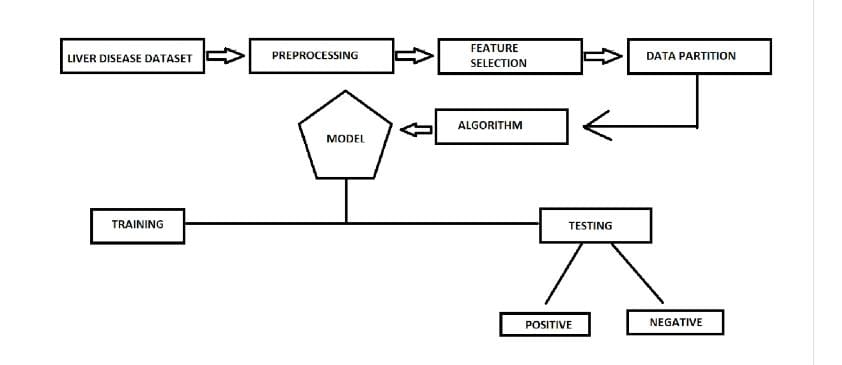
**5.1 DATA FLOW DIAGRAMS**



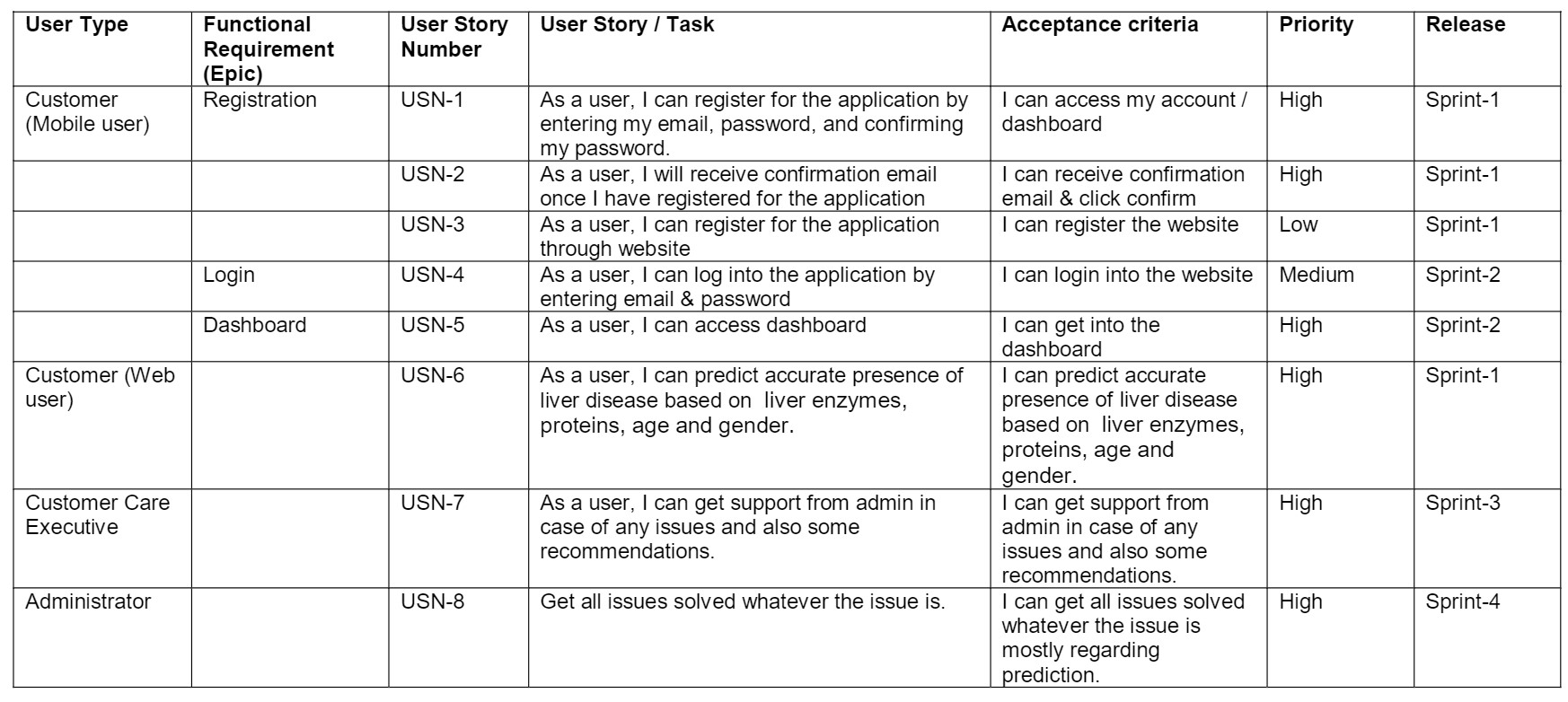
**5.2 SOLUTION ARCHITECTURE**



**TECHNICAL ARCHITECTURE:**



**5.3 USER STORIES**



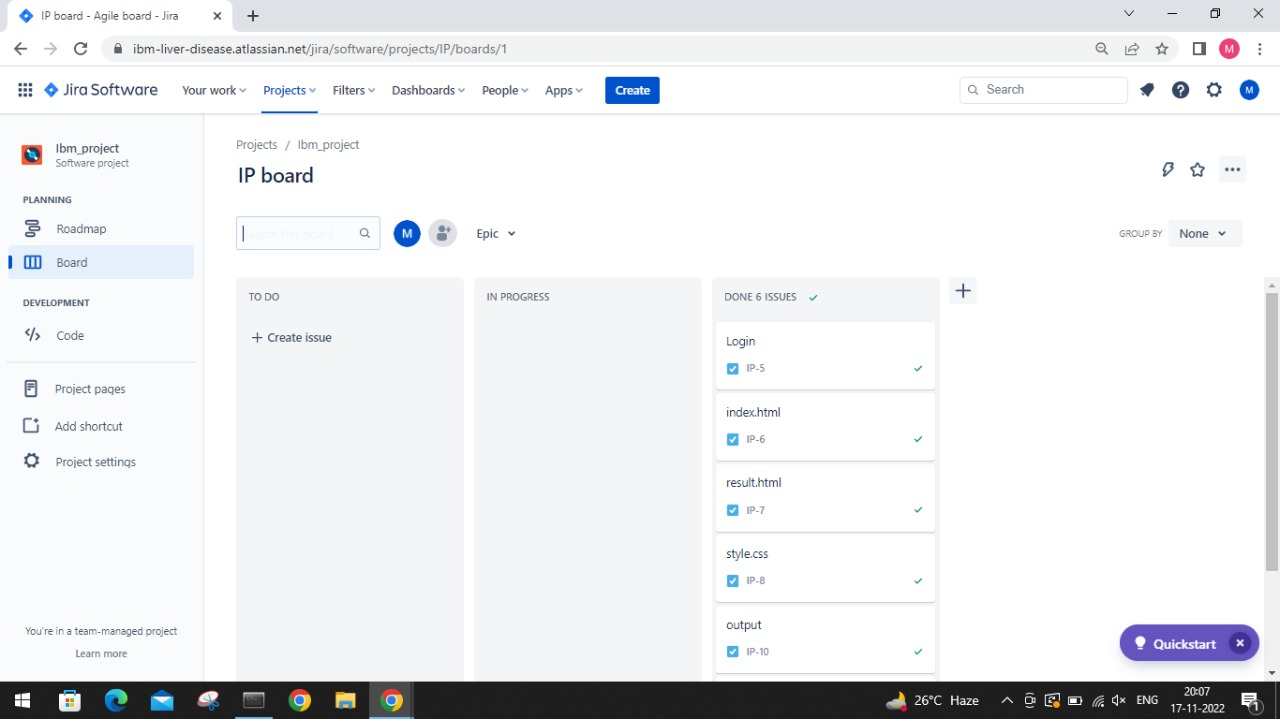
**6.1PROJECT PLANNING & SCHEDULING**

| **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 |

* 1. **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 |

* 1. **REPORTS FROM JIRA**



|  |  |
| --- | --- |
|  |  |
|  |
|  |
|  |
|  |
|  |
|  |  |
|  |  |
|  |  |
|  |  |

**7.CODING AND SOLUTIONING**

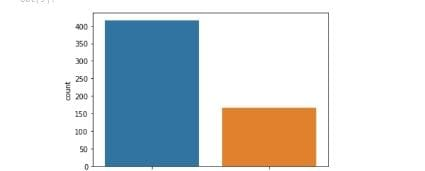
data**.**head()

#to display top five rows of the data



**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:

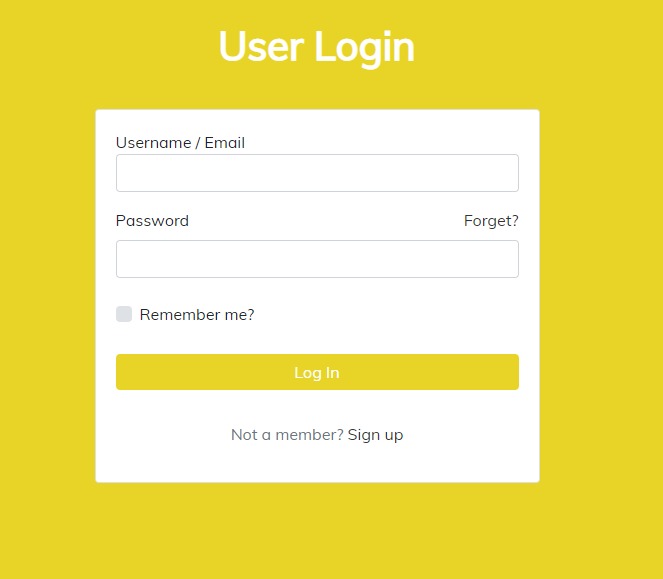




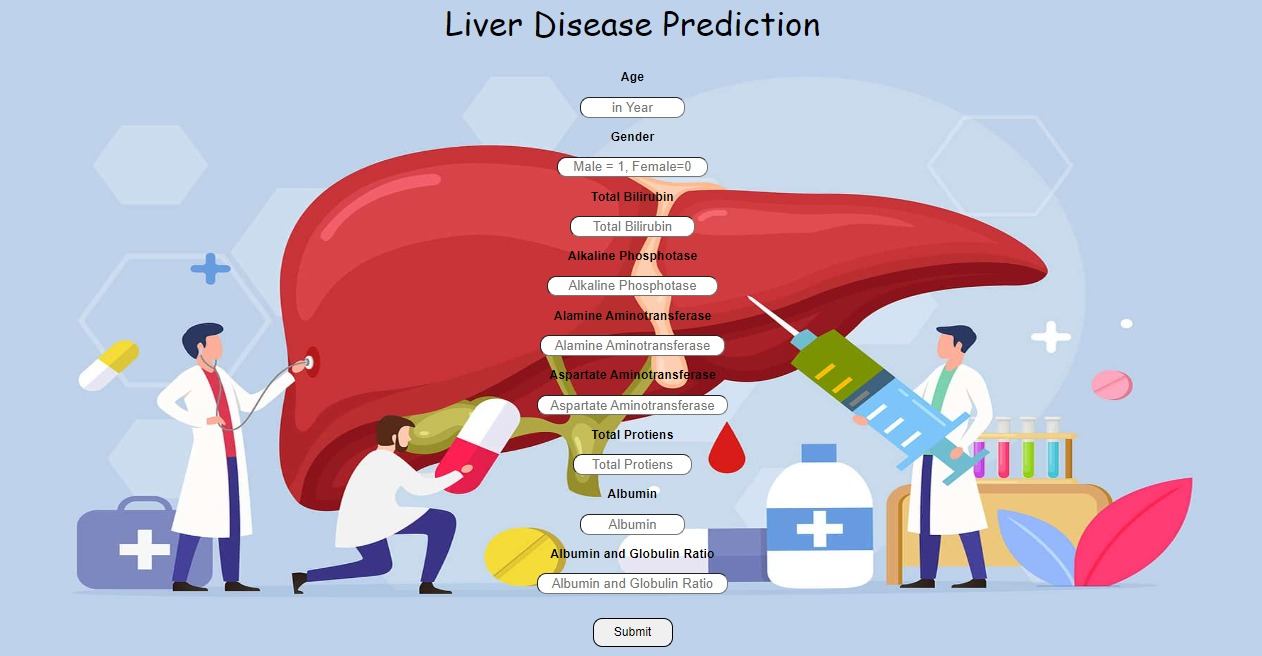


**RESULTS**

Website login page:



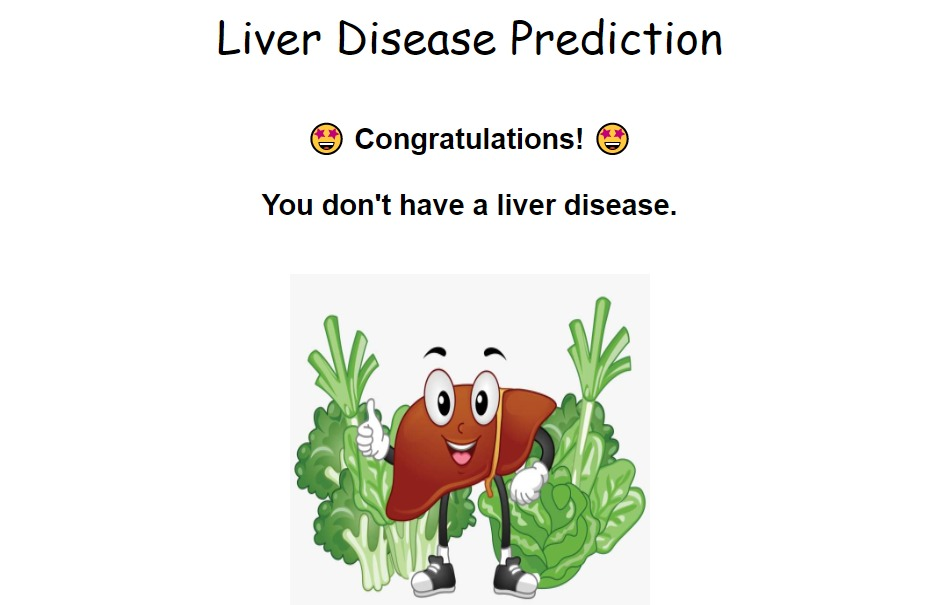
Liver disease prediction



Liver disease predicted



No liver disease predicted



**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.

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**

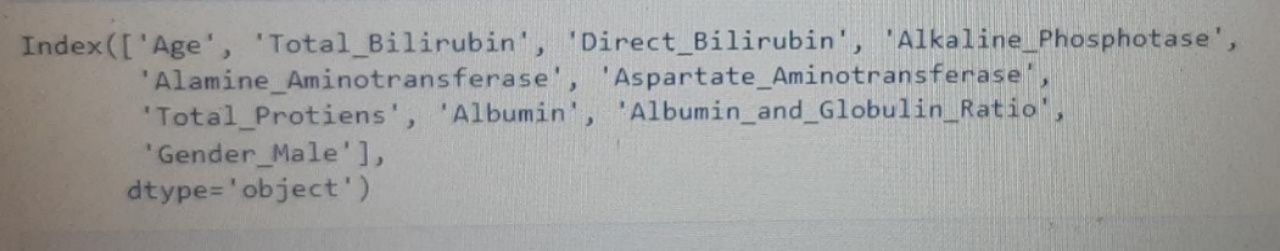
|  |  |
| --- | --- |
|  | Source code:  **import** numpy **as** np  **import** pandas **as** pd  **import** matplotlib.pyplot **as** plt  **%matplotlib** inline  #Importing libraries  data **=** pd**.**read\_csv('indian\_liver\_patient.csv')  #reading CSV file using pandas  data**.**head()  #to display top five rows of the data    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.    data**.**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['Albumin\_and\_Globulin\_Ratio'] **=** data['Albumin\_and\_Globulin\_Ratio']**.**fillna(0.947064)  data**.**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 0  Dataset 0  dtype: int64  **import** seaborn **as** sns  sns**.**countplot('Dataset', data **=** data)    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)    sns**.**countplot('Gender', data **=** data, hue **=** 'Dataset')    data **=** pd**.**get\_dummies(data, columns **=** ['Gender'], drop\_first **=** **True**)  data**.**head()    plt**.**figure(figsize **=** (20,20))  sns**.**heatmap(data**.**corr(), annot **=** **True**)  sns**.**pairplot(data, hue **=** 'Dataset')  data**.**corr() |

*# 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



**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>

