Statistical Machine Learning Approaches to LiverDisease Prediction

IBM-Project-9305-1658993012

NALAIYA THIRAN PROJECT BASED LEARNING ON PROFESSIONAL READLINESSFOR INNOVATION, EMPLOYNMENT AND ENTERPRENEURSHIP

A PROJECT REPORT

Submitted By

ESWARAN S [921319205030]

HARI KISHORE B [921319205037]

KAMESH P [921319205054]

KARMEGAM R[921319205055]

TEAM ID: PNT2022TMID05432

INDUSTRY MENTOR: Nidhi

FACULTY MENTOR : N.Pandeeswari

EVALUATOR : Mrs.Nithya

PSNA College Of Engineering and Technology Dindigul
624622

Project Report

1. **INTRODUCTION**
    a. Project Overview
    b. Purpose
2. **LITERATURE SURVEY**
    a. Existing problem
    b. References
    c. Problem Statement Definition
3. **IDEATION & PROPOSED SOLUTION**
    a. Empathy Map Canvas
    b. Ideation & Brainstorming
    c. Proposed Solution
    d. Problem Solution fit
4. **REQUIREMENT ANALYSIS**
    a. Functional requirement
    b. Non-Functional requirements
5. **PROJECT DESIGN**
    a. Data Flow Diagrams
    b. Solution & Technical Architecture
    c. User Stories
6. **PROJECT PLANNING & SCHEDULING**
    a. Sprint Planning & Estimation
    b. Sprint Delivery Schedule
    c. Reports from JIRA
7. **CODING & SOLUTIONING (Explain the features added in the project along with code)**
    a. Feature 1

1. INTRODUCTION

a. Project overview:

Liver is the largest internal organ in the human body,it is essential for digesting food and releasing the toxic element of the body and plays a major role in metabolism and serving several vital functions. The liver is the largest glandular organ of the body. It weighs about 3 lb (1.36 kg) .The liver's main job is to strain the blood coming from the digestive tract, before passing it to the rest of the body. The liver also detoxifies chemicals and metabolizes drugs. As it does so, the liver hides bile that ends up back in the intestines. The liver also makes proteins important for blood clotting and other functions.The liver supports almost every organ in the body and is vital for our survival. Liver disease may not cause any symptoms at earlier stage or the symptoms may be vague, like weakness and loss of energy. Symptoms partly depend on
the type and the extent of liver disease. Liver diseases are diagnosed based on the liver functionaltest

b. Purpose

With a growing trend of sedentary and lack of physical activities, diseases related to liver have become a common encounter nowadays. In rural areas the intensity is still manageable, but in urban areas, and especially metropolitan areas the liver disease is a very common sighting nowadays. Liver diseases cause millions of deaths every year. Viral hepatitis alone causes 1.34 million deaths every year. Problems with liver patients are not easily discovered in an early stage as it will be functioning normally even when it is partially damaged. An early diagnosis of liver problems will increase patient’s survival rate. Liver failures are at high rate of risk among Indians. It is expected that by 2025 India may become the World Capital for

Liver Diseases. The widespread occurrence of liver infection in India is contributed due to deskbound lifestyle, increased alcohol consumption and smoking. There are about 100 types of liver infections. With such alarming figures, it is necessary to have a concern towards tackling these diseases. Afterall, we cannot expect a developed and prosperous nation, with unhealthy youths. In this project we have taken UCI ILPD Dataset which contains 10 variables that are age, gender, total Bilirubin, direct Bilirubin, total proteins, albumin, A/G ratio, SGPT, SGOT and Alkphos and contains 415 as liver disease patients and 167 as non-liver disease patients. As we got through the next parts of this paper, we will explain what process as taken place for the selection of best model and building necessary system for the prediction of liver disease.

The major outcomes that can be expected through this project are:

1. Increased convenience for predicting a liver disease
2. Reduction in number of deaths due to liver diseases

More accurate diagnosis of liver disease by the doctors.

2. LITERATURE SURVEY

a. Existing problem

Medical diagnoses have important implications for improving patient care, research, and policy. For a medical diagnosis, health professionals use different kinds of pathological methods tomake decisions on medical reports in terms of the patients' medical conditions. Recently, clinicians have been actively engaged in improving medical diagnoses. The use of artificial intelligence and machine learning in combination with clinical findings has further improved disease detection. In the modern era, with the advantage of computers and technologies, one can collect data and visualize many hidden outcomes such as dealing with missing data in medical research. Statistical machine learning algorithms based on specific problems can assist one to make decisions.To detect disease, healthcare professionals need to collect samples from patients which can cost both time and money. Often, more than one kind of test or many samples are needed from the patient to accumulate all the necessary information for a better diagnosis.Using machine learning algorithms to predict disease is made possible by increasing access to hidden attributes in medical data sets. Various kinds of data sets, such as blood panels with liver function tests, histologically stained slide images, and the presence of specific molecular markers in blood or tissue samples, have been used to train classifier algorithms to predict liver disease with good accuracy. The ML methods described in previous studies have been evaluated for accuracy by a combination of confusion matrix,receiver operating characteristic under area under curve, and k-fold cross-validation.Chronic liver disease is detected by clinicians who are well trained in identifying significant observations and classifying them as normal or abnormal using background information and other context clues. ML algorithms can be trained to detect the possibility of liver disease in a similar way to assist healthcare workers. Using the correlation of each variable with the risk of liver disease to train the model, ML methods were able to identify which blood donors were healthy and which had liver disease with high accuracy.

b. References

1.Liver Disease Prediction System using Machine Learning Techniques, Rakshith D B, Mrigank Srivastava,Ashwani Kumar,Gururaj S P, International Journal of Engineering Research & Technology (IJERT) ,Vol. 10 Issue 06, June-2021

2.A Prediction Model of Detecting Liver Diseases in Patients using Logistic Regression of Machine Learning PSM Keerthana, Nimish Phalinkar, Riya Mehere, Koppula Bhanu Prakash Reddy, Nidhi Lal, INTERNATIONALCONFERENCEON INNOVATIVE COMPUTING AND COMMUNICATION (ICICC-2020).

3. Prediction of Liver Disease Using Machine Learning Algorithm and Genetic Algorithm B.Poonguzharselvi , Mohammad Mahaboob Ali Ashraf , Vadlamani V S S Subhash , S.Karunakaran ,Annals of R.S.C.B, 2021: Volume 25: Issue 4.

4. Machine Learning Techniques in Analysis and Prediction of Liver Disease Dr. Dattatreya P Mankame , Harshitha R , Navya N C , Nitin Ravichander, July 2021| IJIRT | Volume 8 Issue 2 | ISSN: 2349-6002.

5. A Comparative Study On Liver Disease Prediction Using Supervised Machine Learning Algorithms A.K.M Sazzadur Rahman, F. M. Javed Mehedi Shamrat, Zarrin Tasnim, INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 8, ISSUE 11, NOVEMBER 2019 ISSN 2277-8616.

c. Problem Statement Defifinition

The main objective of this project is to analyse the parameters of various classification algorithmsand compare their precision, recall, F1 score and accuracies so as to find out the best classifier for determining the liver disease. Here we are building a model by applying various machine learning
algorithms to find the best accurate model and integrate it to a flask-based web application. User can predict the disease by entering the values in the web

application. So many statistical and machine learning approaches (e.g., simulation modelling, classification, and inference) have been used by researchers and lab technicians for better prediction. The clinical results are more data driven than model-dependent.

Problem Statement (PS)	**I am (Customer)**	**I am trying to**	**But**	**Because**	**Which makes me feel**
PS-1	Patient	Check if I have Liver disease	The implementation of this idea is difficult	There are variations in accuracies, which are unavoidable	depressed
PS-2	Patient	Simply experiment with the app	The implementation of this project is difficult	There are variations in accuracies, which are unavoidable	Sad

3. IDEATION & PROPOSED SOLUTION

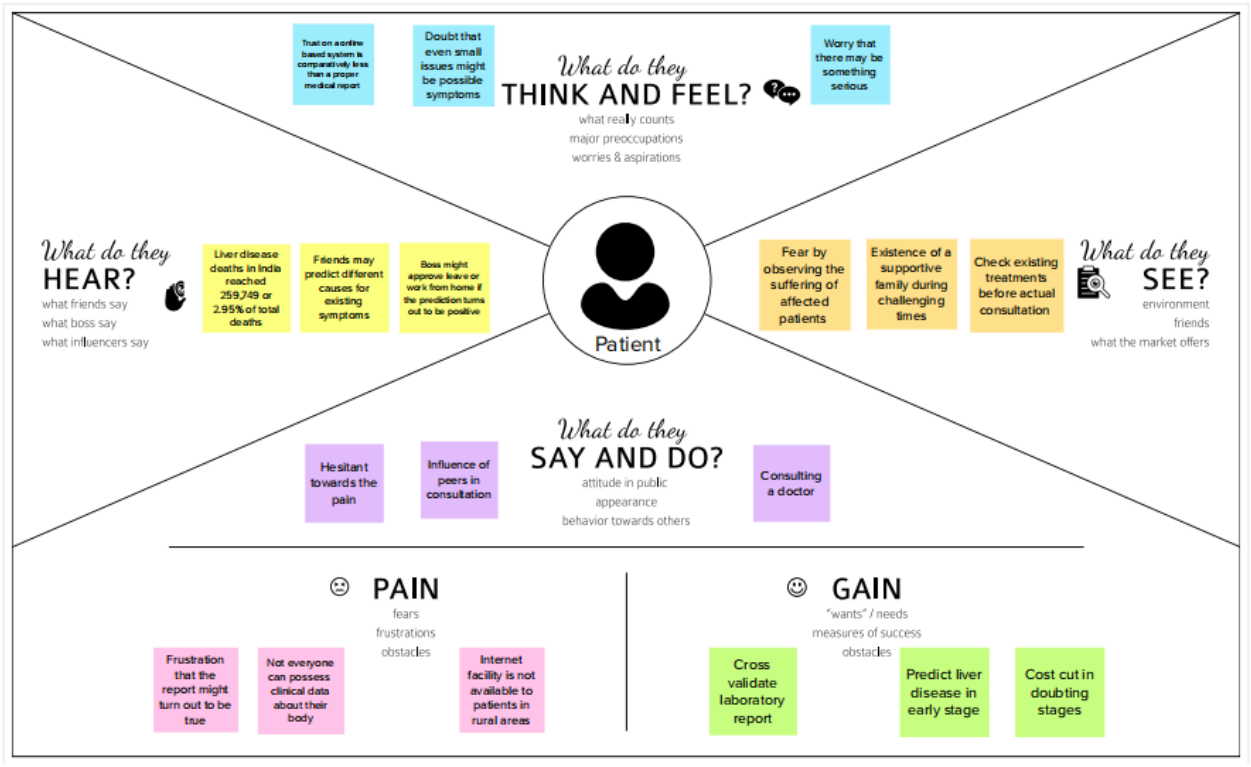
a. Empathy Map Canvas

i. An empathy map is a collaborative tool teams can use to gain a deeper insight into their customers. Much like a user persona, an empathy map can represent a group of users, such as a customer segment. The empathy map was originally created by Dave Gray and has gained much popularity within the agile community.

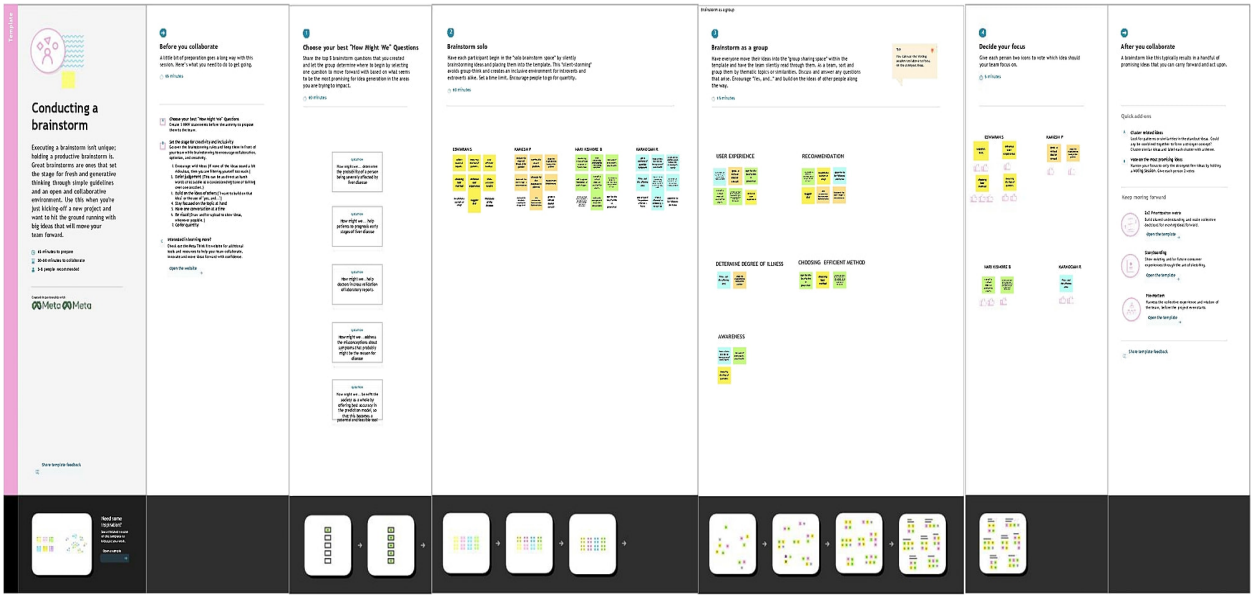
ii. An empathy map is an effective visualization template that helps analyse the behaviour and emotions of customers and users. Empathy maps not only detect the behaviours but highlight possible mediums for brands to communicate with their customers in a better way

iii. Empathy maps can also be used to collect data directly from the users. Used alongside user interviews, survey answers, etc., you can also have a user fill in an empathy map themselves. This often reveals aspects of the user that may have remained unsaid or not thought of.

iv. Each of the four quadrants comprises a category that helps us delve into the mind of the user. The four empathy map quadrants look at what the user says, thinks, feels, and does.



b.Ideation & Brainstorming

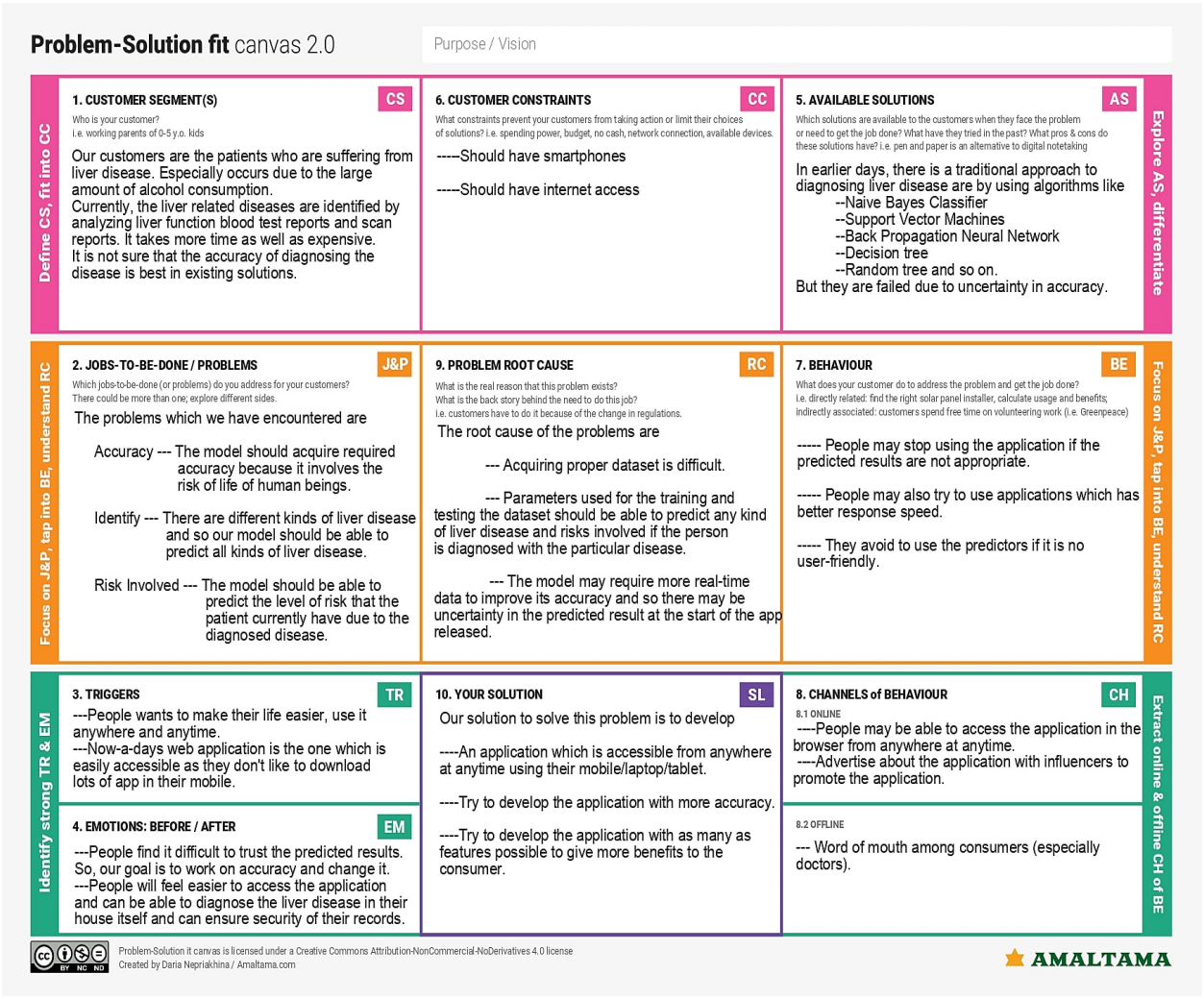


c. Proposed Solution

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 fast and accurate and to diagnose the patients in early stage .
2.	Idea / Solution description	Machine learning model which uses statistical data to predict the liverdisease of the patients.
3.	Novelty / 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 them to predict the likeliness of the liver disease
4.	Social Impact / CustomerSatisfaction	● Capable of predicting the liverdisease in early stage ● Works accurately and precisely to 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 the liver disease. ● The user can be able to get consulting with doctors
6.	Scalability of the Solution	● Can be extended to predict many classification of diseases in early stage ● This can be integrated to with any hospitals and health sectors to get patient records securely through APIs

d. Problem Solution Fit



4. REQUIREMENT ANALYSIS

a. Functional Requirements

Following are the functional requirements of the proposed solution.

FR No.	Functional Requirement (Epic)	Sub Requirement (Story / Sub-Task)
FR-1	User Registration	New user account can be created through web application.
FR-2	User Confirmation	The system gives an approval message after the user account is activated.
FR-3	User medication data	Data should be fed to the dashboard text fields in the application.
FR-4	Database Management	User data will be saved in the database and will be used for future reference.
FR-5	Reporting	Predicting liver disease using given data and generating the medical report for future use.

b. Non-Functional Requirement

Following are the non-functional requirements of the proposed solution.

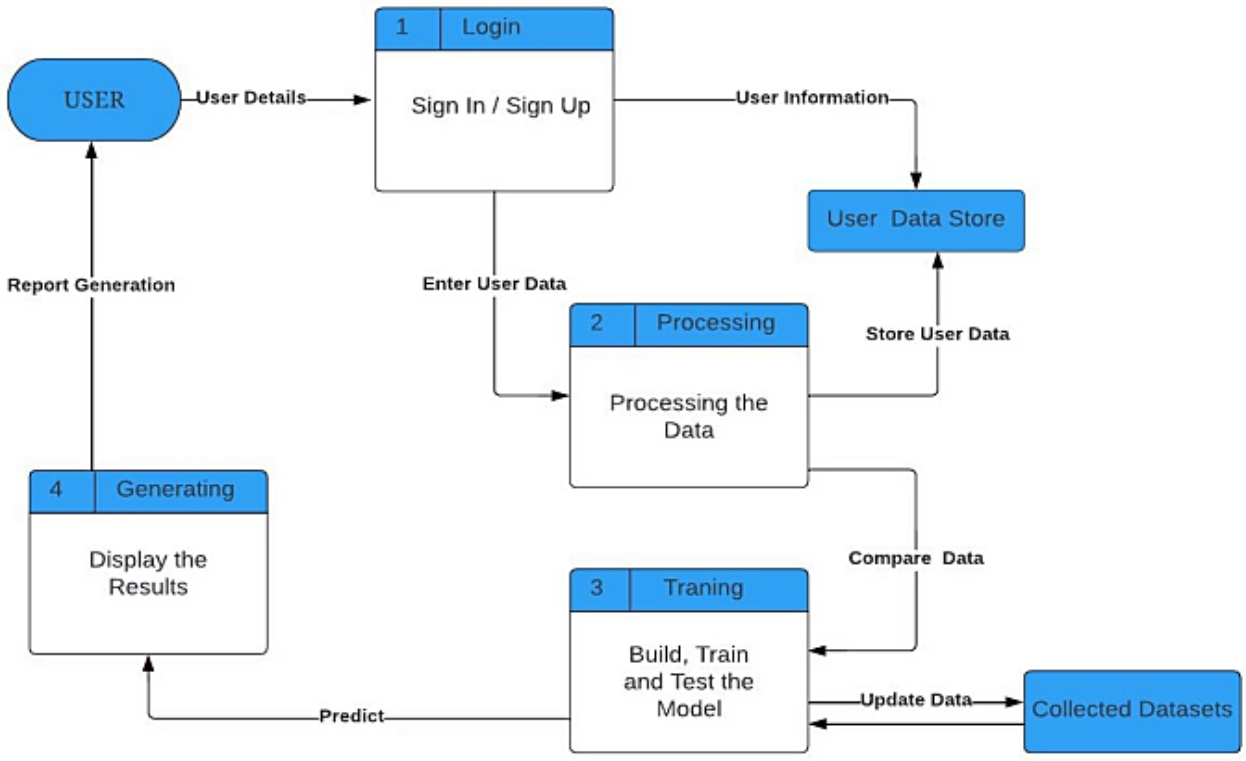
FR No.	Non-Functional Requirement	Description
NFR-1	**Usability**	The whole system can be accessed through web application. Hence it is very

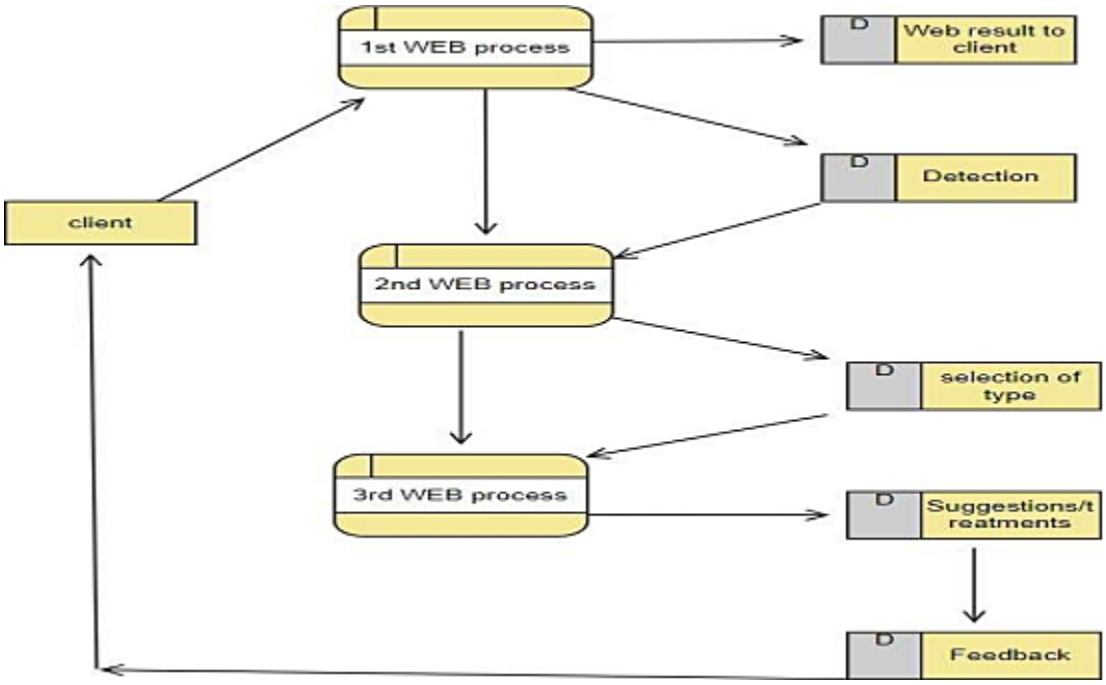
		easy to use.
NFR-2	**Security**	Security requirements ensure that the software is protected from unauthorized access to the system and its stored in data.
NFR-3	**Reliability**	Support vector machine (SVM), Random Forest algorithm and KNN algorithm have been employed for developing liver disease risk prediction model and obtained the accuracy.
NFR-4	**Performance**	Application effectively compares user given parameters with the required dataset. Hence performance would be considerably good.
NFR-5	**Availability**	It is gauged by period that system's functionality & services are available for use with all operations.
NFR-6	**Scalability**	Application can be used in any kind of operating system either in small or large OS so the scalability is very high.

5. PROJECT DESIGN

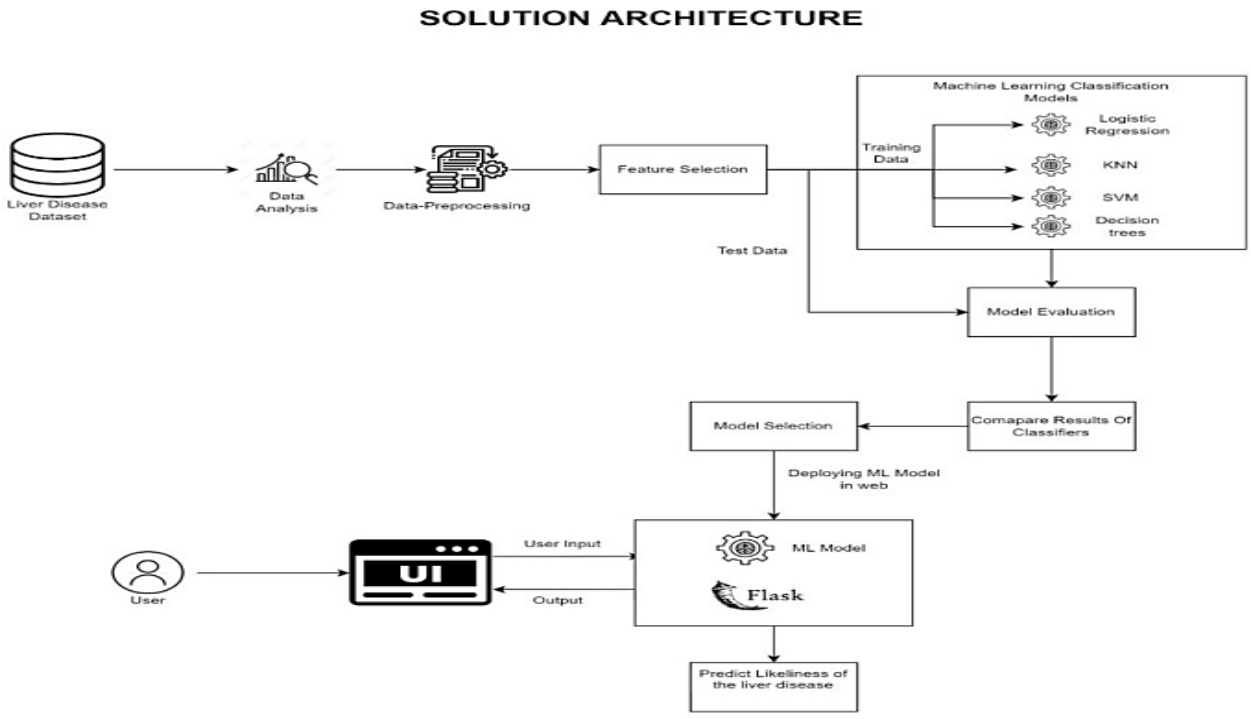
a. Data Flow Diagrams

A Data Flow Diagram (DFD) is a traditional visual representation of the information flows within a system. A neat and clear DFD can depict the right amount of the system requirement graphically. It shows how data enters and leaves the system, what changes the information, and where data is stored.

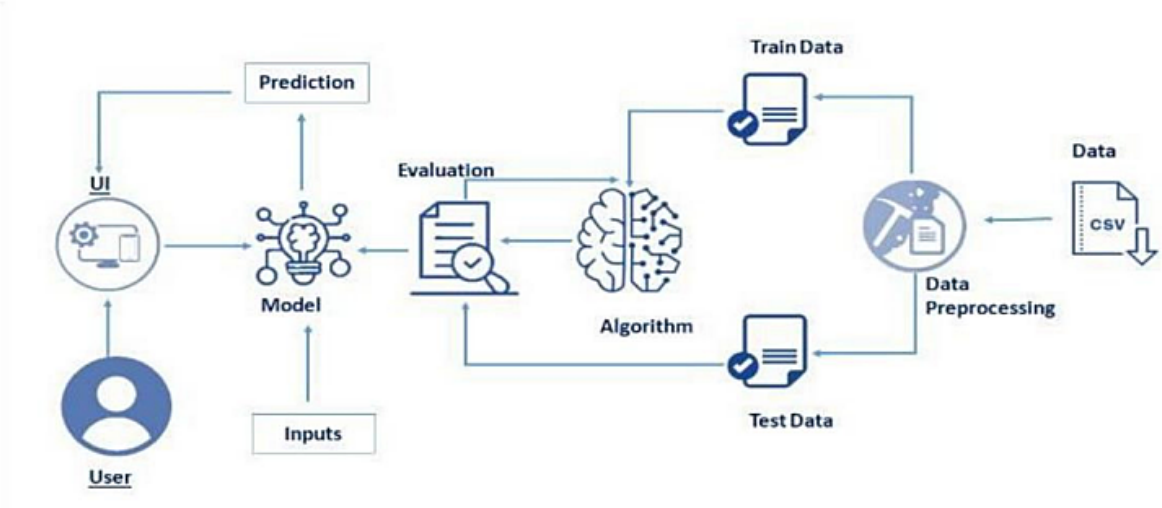




b.Solution and Technical Architecture



Technical Architecture:



c. User Stories

Use the below template to list all the user stories for the product

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 canregister forthe application by enteringmy 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
	Login	USN-3	As a user, I can log into the application by entering email	I can access the dashboard	High	Sprint-1

			& password			

	Dashboard	USN-4	As a user, I must enter my details	I can retrieve information anywhere	Medium	Sprint-2
	Dashboard	USN-5	As a user, I must enter my details	I can view various pages	High	Sprint-1
Customer (Web user)	Upload Images	USN-6	As a user, I can upload the image that required for finding whether liver disease is there are not.	Can get result based on the information provided.	Medium	Sprint-3
	Enter data	USN-7	As a user, I can enter the required data from the scanned report.	Can get result based on the information provided	High	Sprint-3
	Report	USN-8	As a user, I can generate the report in PDF format.	Result can be generated in PDF format in user login.	Low	Sprint-4
	Search	USN-9	As a user, I can search for the specialist and best hospital in that respective field.	I can receive information on various doctors and hospitals.	Low	Sprint-4
Administrator	Analyse	USN-10	As an admin, I will analyse the given data	I can analyse the given data.	High	Sprint-2

6. PROJECT PLANNING & SCHEDULING

a. Sprint 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 applicationby enteringmy email, password, and confirming my password.	5	High	Eswaran S
Sprint-1		USN-2	As a user, I will receive confirmation email Once I have registered for the application	5	High	Kamesh P
Sprint-1	Login	USN-3	As a user, I can log into the application by entering email & password	10	High	Karmegam R
Sprint-2	Input Necessary Details	USN-4	As a user, I can give Input Details to Predict Likeliness of Liver Disease.	15	High	Kamesh P

Sprint-2	Data pre-processing	USN-5	Transform raw data into suitable format for prediction.	5	High	Eswaran S
Sprint-3	Prediction of Liver Disease	USN-6	As a user, I can predict Liver Disease Using machine learning model.	15	High	Hari KishoreB
Sprint-3	.	USN-8	As a user, I can get accurate prediction of Liver disease.	5	Medium	Karmegam R
Sprint-4	Review	UNS-8	As a user, I can give feedback of The application	20	High	Eswaran S

b. Sprint Delivery Schedule

Sprint	**Total Story Points**	**Duration**	**Sprint Start Date**	**Sprint End Date(Planned)**	**Story Points Completed(as on PlannedEnd Date)**	**Sprint Release Date(Actual)**
Sprint-1	20	6 Days	24 Oct2022	29 Oct 2022	18	08 Nov 2022

Sprint-2	20	6 Days	31 Oct 2022	05 Nov 2022	17	06 Nov 2022
Sprint-3	20	6 Days	07 Nov 2022	12 Nov 2022	18	08 Nov 2022
Sprint-4	20	6 Days	14 Nov 2022	19 Nov 2022	17	10 Nov 2022

c. Reports from JIRA

JIRA has categorized reports in four levels, which are −

i. Agile
ii. Issue Analysis
iii. Forecast & Management
iv. Others

VELOCITY: SPRINT - 1

Sprint duration = 5 days

Velocity of team = 20 points

Average Velocity (AV) = $\frac{\text{Velocity}}{\text{Sprint duration}}$

$AV = 20/5 = 4$

Average Velocity = 4

VELOCITY: Sprint 1 - 4

Sprint duration = 20 days

Velocity of team = 80 points

Average Velocity (AV) = $\frac{\text{Velocity}}{\text{Sprint duration}}$

$AV = 80/20 = 4$

Total Average Velocity = 4

7. CODING & SOLUTIONING

a.Feature 1

Flaskapp.py

```
from flask import Flask, render_template, request
import numpy as np
import pickle


app = Flask(__name__)
model = pickle.load(open('SVM.pkl', 'rb'))

@app.route('/',methods=['GET'])
def Home():
    return render_template('index.html')

@app.route("/predict", methods=['POST'])
def predict():
    if request.method == 'POST':
        Age = int(request.form['Age'])
        Gender = 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'])


        values =
np.array([[Age,Gender,Total_Bilirubin,Direct_Bilirubin,Alkaline_Phosphotase,Ala
mine_Aminotransferase,Aspartate_Aminotransferase,Total_Protiens,Albumin,Albumin
_and_Globulin_Ratio]])
        prediction = model.predict(values)
```

```
        return render_template('result.html', prediction=prediction)


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

b. Feature 2

Scoringendpoint.py

```
from flask import Flask, render_template, request
import numpy as np
import pickle
import requests
import json

# NOTE: you must manually set API_KEY below using information retrieved from
your IBM Cloud account.
API_KEY = "AluxVbbOjio3fHsuNSWWK-rZXX54vtrvqLSJTj5QTt4I"
token_response = requests.post('https://iam.cloud.ibm.com/identity/token',
data={"apikey":
 API_KEY, "grant_type": 'urn:ibm:params:oauth:grant-type:apikey'})
mltoken = token_response.json()["access_token"]
print("mltoken",mltoken)

header = {'Content-Type': 'application/json', 'Authorization': 'Bearer ' +
mltoken}

app = Flask(__name__)
model = pickle.load(open('SVM.pkl', 'rb'))

@app.route('/',methods=['GET'])
def Home():
    return render_template('index.html')

@app.route("/predict", methods=['POST'])
def predict():
    if request.method == 'POST':
        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'])


        values =
np.array([[Age,Gender,Total_Bilirubin,Direct_Bilirubin,Alkaline_Phosphotase,Ala
mine_Aminotransferase,Aspartate_Aminotransferase,Total_Protiens,Albumin,Albumin
_and_Globulin_Ratio]])
        prediction = model.predict(values)

        return render_template('result.html', prediction=prediction)


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

8. TESTING

Test Cases

- The home page and the result page is tested .It is working wellwithout issues.
- The app was tested for functionality .

- The scoring end point application is slower than the normal flask app.

9. RESULTS

a.Performance metrics

Support Vector Matrix
Confusion Matrix

```
[[115   0]
 [ 60   0]]
```

Accuracy

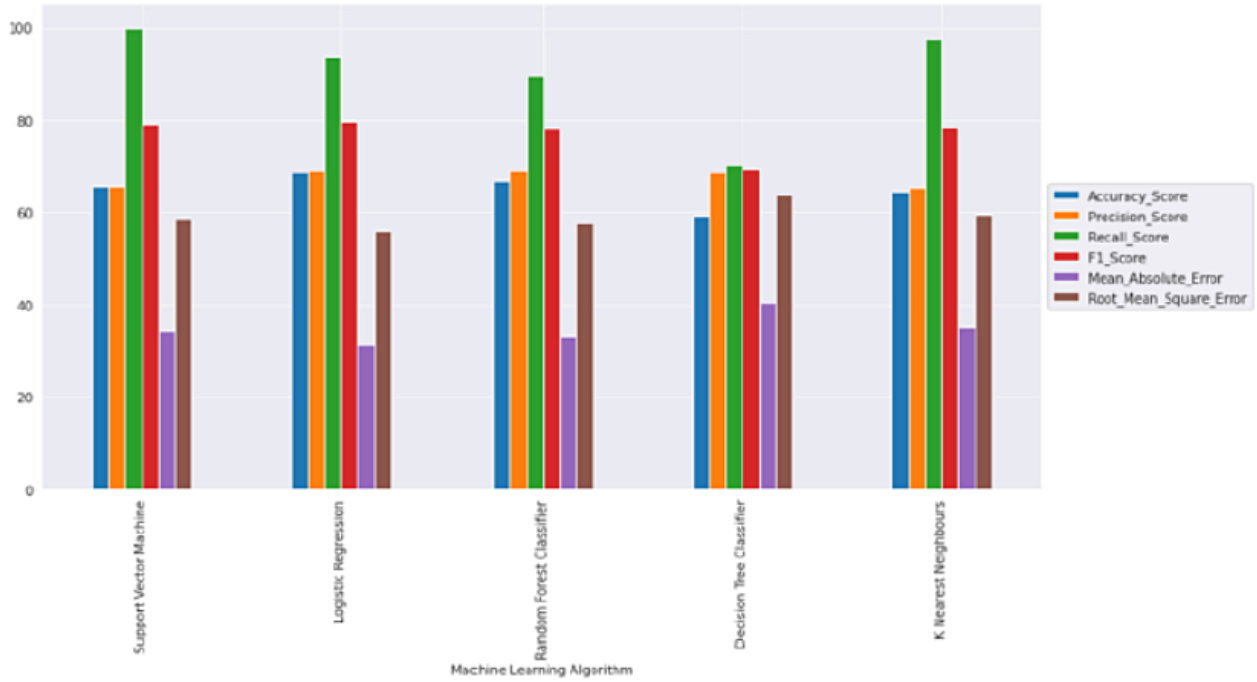
	Machine Learning Algorithm	Accuracy_Score
0	Support Vector Machine	65.714286

Classification Result

	precision	recall	f1-score	support
1	0.66	1.00	0.79	115
2	0.00	0.00	0.00	60
accuracy			0.66	175
macro avg	0.33	0.50	0.40	175
weighted avg	0.43	0.66	0.52	175

Comparision with other matrix

	Machine Learning Algorithm	Accuracy_Score	Precision_Score	Recall_Score	F1_Score	Mean_Absolute_Error	Root_Mean_Square_Error
0	Support Vector Machine	65.714286	65.714286	100.000000	79.310345	34.285714	58.554004
1	Logistic Regression	68.571429	69.230769	93.913043	79.704797	31.428571	56.061191
2	Random Forest Classifier	66.857143	69.127517	89.565217	78.030303	33.142857	57.569833
3	Decision Tree Classifier	59.428571	68.644068	70.434783	69.527897	40.571429	63.695705
4	K Nearest Neighbours	64.571429	65.497076	97.391304	78.321678	35.428571	59.521905



Hypertune the model using GrideSearch CV

```
from sklearn.svm import SVC
```

```
model=SVC()
```

```
param_grid = {'C': [0.1,1, 10,50,100,200,1000], 'gamma': [1,0.1,0.01,0.001,0.0001], 'kernel': ['rbf']}
from sklearn.model_selection import GridSearchCV
grid = GridSearchCV(SVC(),param_grid,refit=True,verbose=3)
grid.fit(X_train,y_train)
```

```
Fitting 5 folds for each of 35 candidates, totalling 175 fits
[CV 1/5] END ........C=0.1, gamma=1, kernel=rbf;, score=0.707 total time=   0.0s
[CV 2/5] END ........C=0.1, gamma=1, kernel=rbf;, score=0.707 total time=   0.0s
[CV 3/5] END ........C=0.1, gamma=1, kernel=rbf;, score=0.707 total time=   0.0s
[CV 4/5] END ........C=0.1, gamma=1, kernel=rbf;, score=0.716 total time=   0.0s
[CV 5/5] END ........C=0.1, gamma=1, kernel=rbf;, score=0.716 total time=   0.0s
[CV 1/5] END ......C=0.1, gamma=0.1, kernel=rbf;, score=0.707 total time=   0.0s
```

```
GridSearchCV(estimator=SVC(),
             param_grid={'C': [0.1, 1, 10, 50, 100, 200, 1000],
                         'gamma': [1, 0.1, 0.01, 0.001, 0.0001],
                         'kernel': ['rbf']},
             verbose=3)
```

```
]  svm_predictions=grid.predict(X_test)
```

	precision	recall	f1-score	support
0	0.00	0.00	0.00	43
1	0.75	1.00	0.86	132
accuracy			0.75	175
macro avg	0.38	0.50	0.43	175
weighted avg	0.57	0.75	0.65	175

10. ADVANTAGES AND DISADVANTAGES

Advantages :

- This helps in early diagnosis of liver disease.
- It makes the process simple and easier.
- We are able to monitor the patient in an effective manner.
- We can also help patients who are not in a condition to directly consult a doctor.

Disadvantages:

- There are inaccuracies which cannot be avoided.
- For people who have difficulty in accessing internet or those who do not have electronic gadgets , this is not a feasible solution.

11. CONCLUSION

Classification is the major data mining technique which is primarily used in healthcare sectors for medical diagnosis and predicting diseases. This project work used classification algorithms namely Support Vector Machine (SVM) for liver disease prediction.

Comparisons of different algorithms are done and it is based on the performance factors classification accuracy and execution time. From the experimental results, we concluded that SVM classifier is considered as the best.

12. FUTURE SCOPE

In future , we can add extra features for training the model in our proposed system .We can use different machine learning algorithms and test the performance of them based on .If we find a better algorithm with high accuracy , we can use it.Then we can add extra suggestions like reminders for medicines etc.

13. APPENDIX

Github link: https://github.com/IBM-EPBL/IBM-Project-9305-1658993012

Demolink:https://drive.google.com/drive/folders/14JMU9gV8ndsLGOyTQbUQeAoRuhsinUB2?usp=share_link nYw/view?usp=sharing