STATISTICAL MACHINE LEARNING APPROACHES TO LIVER DISEASE PREDICTION

Team ID : PNT2022TMID14267

Team Members :

➢ Hemamalini B
➢ C Sri Vaishnavi
➢D Sowmya Sen
➢ Harsha Varthini S

1. INTRODUCTION

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 proccess as taken place for the selection of best model and building necccessary sytem for the prediction of liver disease. The major outcomes that can be expected through this project are:

➢ Increased convenience for predicting a liver disease
➢ Reduction in number of deaths due to liver diseases
➢ More accurate diagnosis of liver disease by the doctors

1.1 Project Overview

Liver Disease is a prominent Disease other than heart attack, which is taking a lot of lives. Since most of the time liver disease is detected at the later stage leading to death. Number of liver patients is increasing because of several reasons like over consumption of alcohol, breathing in injurious gas, consuming polluted water and so on which can affect health parameter. Using a machine learning prediction models, liver diseases can be predicted using those parameters in early stages. In this work to build the machine-learning models, Indian Liver Patient Dataset (ILPD) hosted at UCI.edu is used, which is based on Indian patient and Naive Bayes algorithm is used to predict the disease with different prepossing techniques. Dataset is checked for

skewness, outliers and imbalance using univariate and bivariate analysis and then suitable algorithms used to remove outliers and various oversampling and undersampling techniques are used to balance the datal. Further refinement of model is done through hyperparameter tunning using grid search and feature selection. The Final model provides 100% accuracy and also good score across different metrics.

1.2 Purpose

The purpose of this project is used to predict the liver disease using Naive Bayes Algorithm and this is used to predict the liver disease earlier and reduce the death rate .

2. Literature Survey

1. A COMPARATIVE STUDY ON LIVER DISEASE PREDICTION USING SUPERVISED MACHINE LEARNING ALGORITHMS A.K.M Sazzadur Rahman, F. M. Javed Mehedi Shamrat, Zarrin Tasnim, Joy Roy, Syed Akhter Hossain Chronic Liver Disease is the leading cause of global death that impacts the massive quantity of humans around the world. This disease diagnosis is very costly and complicated. Therefore, this paper evaluates the performance of different Machine Learning algorithms in order to reduce the high cost of chronic liver disease diagnosis by prediction. Six machine learning techniques have been applied including Logistic Regression, K Nearest Neighbours, Decision Tree, Support Vector Machine, Naïve Bayes, and Random Forest. The performance was evaluated on different measurement techniques such as accuracy, precision, recall, f-1 score, and specificity and the result was that LR achieved the highest accuracy. --ResearchGate – 2019

2. MACHINE LEARNING-BASED LIVER DISEASE DIAGNOSIS:
A SYSTEMATIC REVIEW Rayyan Azam Khan, Yigang Luo, Fang Xiang Wu
This paper mainly focuses on the computer-aided diagnosis of hepatic lesions in view of diffuse- and focal liver disorders. This is based on three image acquisition modalities: ultrasonography, computed tomography, and magnetic resonance imaging. Insightful analysis is presented for each preliminary step, particularly preprocessing, attribute analysis, and classification techniques to accomplish clinical diagnostic tasks. In preprocessing denoising, deblurring, and segmentation methods are used. Denoising is mainly performed with nonlinear models. --ScienceDirect – 2022

3. DIAGNOSING OF LIVER DISEASE PREDICTION IN PATIENTS USING COMBINED MACHINE LEARNING MODELS
Chokka Anuradha, D Swapna, Balamuralikrishnan Thati IEEE
In the human body one of the most important organs is the liver. If the regular functionality of the liver is disturbed then this condition is called disease-affected liver. Therefore, an early stage of disease detection is more important which helps in disease prevention at starting stage with small medications. But it is too difficult to identify Liver disease at the early stages because symptoms are very less at the starting stage. Lab results with physical examination are involved in the Traditional methods. This paper aims to represent a Diagnosing for Liver disease prediction in Patients using Combined Machine Learning Models. Optimized three machine learning algorithms are used for the accurate diagnosis of liver disease and they are Artificial Neural Networks (ANN), Decision Trees, and K-Nearest Neighbours (KNN). With the help of these algorithms, given data is classified and results are produced. The future data is predicted with the help of past and present data. The accuracy results are produced by comparing three classification algorithms

. 4. STATISTICAL MACHINE LEARNING APPROACHES TO LIVER DISEASE PREDICTION
Fahad Mostafa, Easin Hasan, Morgan Williamson, Hafiz Khan M
L algorithms are trained to detect the possibility of liver disease to assist healthcare workers. Correlation of chosen variable with the risk of liver disease is performed to train the model. ML methods were able to

identify the liver disease with high accuracy. The PCA results showed five important factors for liver disease diagnosis: AST, ALT, GGT, BIL, and ALP. In a real situation, a clinician can strongly suspect liver disease using only these five variables, as they are very descriptive for liver function. The ratio of ALT and AS denotes the cause of a liver injury. GGT and ALP increase in circulation with the severity of a liver injury. Additionally, the injury proximity to the bile duct is determined by the concentration of ALP. This study shows several machine learning approaches with PCA, which outperformed the classification. Among three ML classification methods, the performance of SVM and RF is better than ANN. --MDPI - 2021

5. LIVER DISEASE PREDICTION SYSTEM USING MACHINE LEARNING TECHNIQUES
Rakshith D B, Mrigank Srivastava, Ashwani Kumar, Gururaj S P
In this paper risk of liver disease for a person is predicted based on the blood test report results of the user. With the dataset used for this project, 100 % accuracy is obtained for SVM model. The data pre processing was done using Jupyter Notebook and Desktop Application was Implemented using Sypder IDE.The programming language which was used is python and machine learning Sklearn was used to build the model using classification algorithm like KNN,SVM,Naive Bayes and ANN. --IJERT – 2021

6. STATISTICAL MACHINE LEARNING APPROACHES TO LIVER DISEASE PREDICTION
International Journal of Scientific Research and Engineering Development – 2022
This study attempts to find an appropriate machine learning algorithm that can determine whether a person has liver disease or not given a dataset containing biological and diagnostic data of 583 Indian patients. Using certain characteristics such as total bilirubin, direct bilirubin, alkaline phosphatase, total protein, albumin, and globulin, this software can determine whether a patient has liver disease or not. --Robin Biju

2.1 Existing Problem

The main Problem is doctors cannot diagnose on the basis of variation in the test result .In this Application ,by using patient records that includes blood test report results,we will determine which patient has liver disease and which ones do not in an accurate and faster way.

2.2 References

1.A.K.M Sazzadur Rahman, F. M. Javed Mehedi Shamrat, Zarrin Tasnim, Joy Roy, Syed Akhter Hossai
2. MACHINE LEARNING-BASED LIVER DISEASE DIAGNOSIS A SYSTEMATIC REVIEW Rayyan Azam Khan, Yigang Luo, Fang Xiang Wu
3.. DIAGNOSING OF LIVER DISEASE PREDICTION IN PATIENTS USING COMBINED MACHINE LEARNING MODELS Chokka Anuradha, D Swapna, Balamuralikrishnan Thati
4.STATISTICAL MACHINE LEARNING APPROACHES TO LIVER DISEASE PREDICTION Fahad Mostafa, Easin Hasan, Morgan Williamson, Hafiz Khan –-MDPI - 2021
5. LIVER DISEASE PREDICTION SYSTEM USING MACHINE LEARNING TECHNIQUES Rakshith D B, Mrigank Srivastava, Ashwani Kumar, Gururaj S P –---IJERT – 2021
6.. STATISTICAL MACHINE LEARNING APPROACHES TO LIVER DISEASE PREDICTION International Journal of Scientific Research and Engineering Development – 2022—--Robin Biju

2.3 Problem Statement Definition

The main problem is doctor cannot diagnosis on the basis of variations in test results.

3. IDEATION AND PROPOSED SOLUTION

3.1. Empathy Map Canvas

3.2. Ideation and Brainstorming

Step 1: Team Gathering, Collaboration and Select the Problem Statement

Step 2: Brainstorm, Idea Listing and Grouping

Step 3: Idea Prioritization

3.3. Proposed Solution

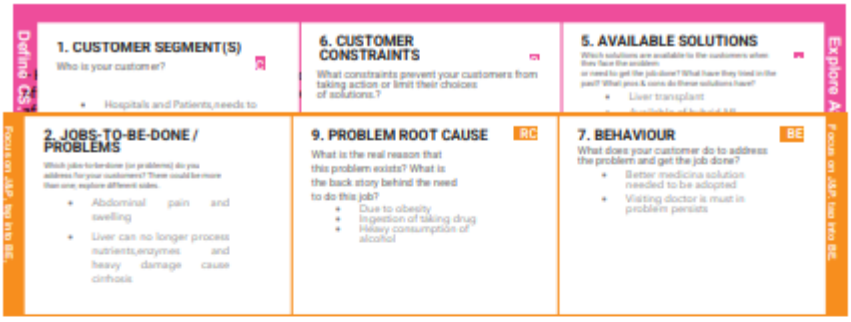
s.no	Parameter Description
1.	Problem Statement (Problem to be solved) Nowadays many peoples are affected by Liver disease because of that so many people are losing their life. Early detection of liver disease can be very helpful in the treatment of the disease to fast recover but it is very difficult to identify the liver disease in early

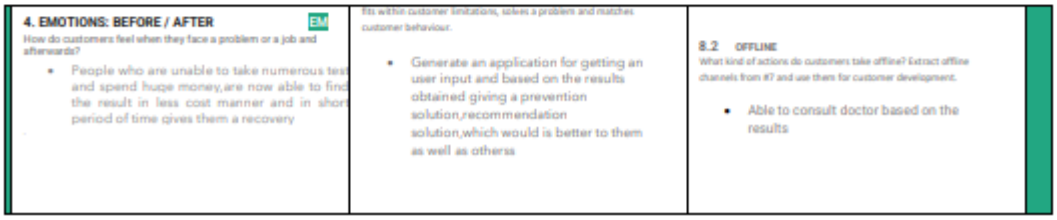
	stages. In some situations, the medical expert is unable to detect the symptom even at an early stage. It is one of the great losses for the patient
2.	Idea / Solution description The software will detect the patients symptoms and it will find out the disease according to that symptom and it will show the result by ML techniques. We will do this by taking the data set of both normal and abnormal liver and we will train the software in that way by detecting the disease according to the symptom.
3.	Novelty / Uniqueness In this software model we're not t also giving some basic precautions like what to do & don't.which makes this model unique

4.	Social Impact / Customer Satisfaction	Many liver disease are left unpredicted.By implementing thi s software liver disease can be diagnosed in early stages which result in the decrease of death r ate
5.	Business Model (Revenue Model)	Currently the global is running with newest technology likewise our project will more helpful to medi cal fields. And the medical institutio n , clinics and hospitals need to pai d for yearly license and get renew yearly to continue the check up using this software

6.	Scalability of the Solution	The software will never show the new types of liver diseases in future because the algorithm and datasets we provided only f or the current liver diseases in case of any new liver diseases founde d in future it will show as a error in output we need to change the algorithm process to show witho ut error in future.

3.4.ProblemSolutionFit





4. REQUIREMENT ANALYSIS

4.1. Functional Requirements

FR No	Functional Requirements(Epic)	Sub Requirement (Story/Sub Task)
FR 1	User Registration	Registration through Gmail
FR 2	User Confirmation	Confirmation via Email
FR 3	Prediction	Liver Disease can be Predicted more Accurat ely by using Support Vector Machine Algorithm

FR 4	Hardware Requirements	2GB RAM(minimum) 100GB HDD(minimum) Intel i3 quad core 1.66G Hz processor(minimum) Internet Connectivity
FR 5	Software Requirements	Windows 7 or higher Python 3.6.0 or higher Visual Studio Code Flask (python platform) HTML Dataset consisting of Liv er Disease Required librarie s Jupiter notebook
FR 6	Other requirements	IBM cloud login Chro me extension features
FR 7	Events	Model needs a capability of retrieving and displaying

		accurate result

4.2. Non-Functional Requirements

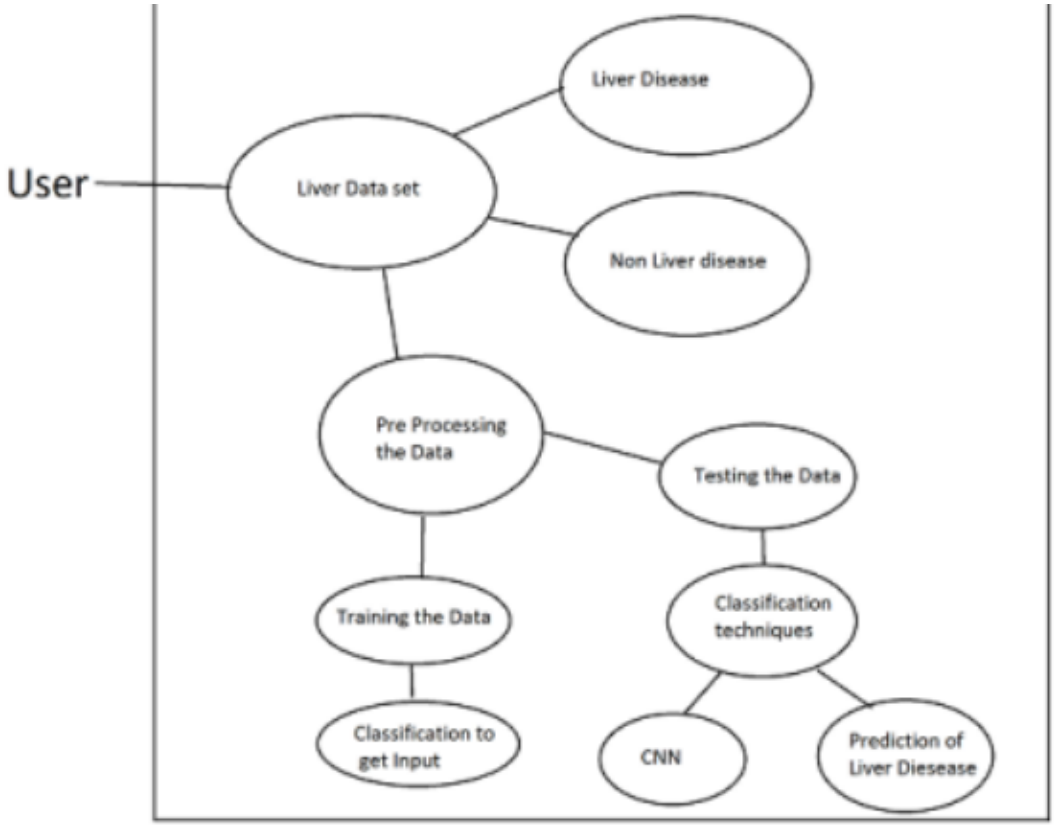
NFR No	Non Functional Requirements	Description
NFR 1	Usability	This system is really use d as it can able to detect Liver Disease .By detecti ng the liver disease early ,death rate is decreased

NFR 2	Security	Assuring all data inside t he system or its part will be Protected
NFR 3	Reliability	This Approach gives mo re accuracy than the existi ng solution
NFR 4	Performance	The effectiveness of the se methods relies on featur e collection, training data, and classification algorithms. It must be processed and execut ed within a fraction of a second using the Machi ne learning algorithm
NFR 5	Availability	It doesn’t have any

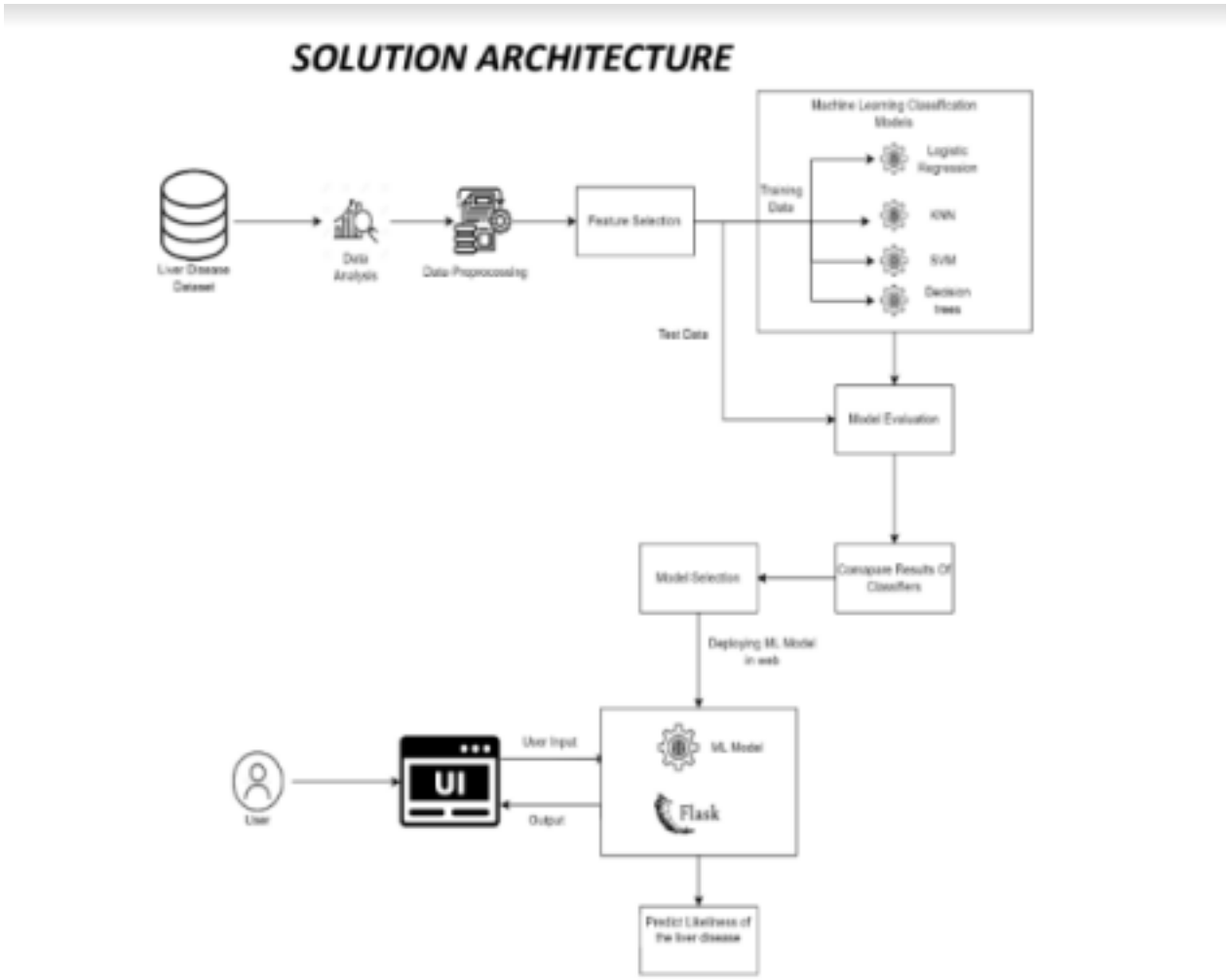
		restrictions , it is availab le for all individual user
NFR 6	Scalability	It is acceptable to fit the m over any place and any resources.

5. PROJECT DESIGN

5.1 Dataflow Diagrams



5.2. Solution Architecture



5.3. Technical Architecture

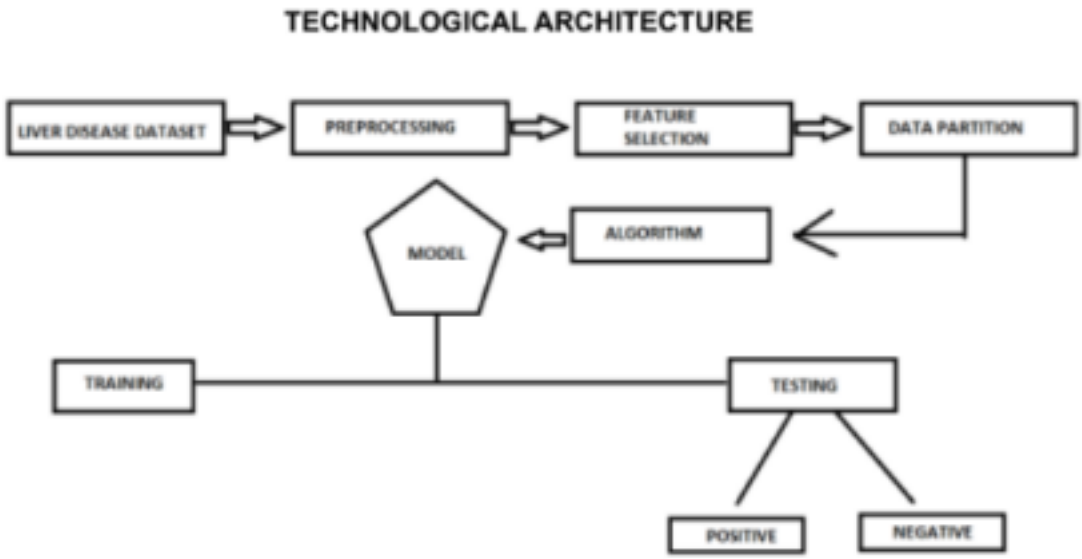


Table-1: Components & Technologies

S.No	Component Description	Technology
1	User Interface How user interact s with application e .g. Web UI, Mobile Ap p, Chatbot etc.	HTML, CSS, JavaScript
2	Application Logic-1 Logic for a proce ss in the application	Python
3	Application Logic-2 Logic for a proce ss in the application	IBM Watson S TT service
4	File Storage Files are stored in cloud	IBM Block Stora ge or Other Storage Service or Local
		Filesystem
5	Machine Learning Model Prediction of Liv er Disease	Support Vector Machine Algorith m

6	Infrastructure (Server / Cloud) IBM Cloud App Configuration is a centralized feature management and configuration service on IBM Cloud	IBM Cloud Foundry, Kubernetes, etc.

Table-2: Applications Characteristics

S.No	Characteristics	Description	Technology
1	Open-Source Frameworks .	There are no open source frameworks in this application	Technology of Opensource framework
2	Security Implementations	Block chain technology is used for Security implementation its private framework protects all data. .	Block chain

3	Scalable Architecture	Users are Provided with medical services online	IBM cloud

4	Availability Available for	Technology used

	everyone , no Restrictions	
5	Performance Predicted Result is more accurate	Support Vector Machine Algori th m

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..	High	Sprint 1
	verify / Login **USN-2**	As a user, I will receive confirmation email once I have registered for the application	High	Sprint 1

	Monitoring **USN-3**	As a user, I can do it from any an monitor the account place process to access .	High	Sprint 2
Customer (Web user)	**Dashboard** **USN-4** **Getting information** **USN-5**	All the login Helpful for process and remindin activities do g ne will be the actions displayed on the dashboard As a user, I I can need to gather collect all the the information Information on from the real scenario	Medium High	Sprint 2 Sprint 1

Custom er Care Executive	**Analysing USN-6**	As a User, I need to analy se the informati on and get into t he decision to predict the disease Helpful to Predict t he disease early	High	Sprint 2
Admini str ator	**Ordering USN-7**	As a user, I would order my officers to hel p the user to predict the disease Iam the higher Authority and I can order the m	High	Sprint 2
		As a user, I will I will try to	High	Sprint 1
	USN-8 observe the account. If a ny prediction of disease go es wrong, I wil I surely take actions to avo id them manage t he situation			

	Obeying Orders **USN- 9**	As a user I ne ed to obey my higher office rs command a nd take measur es mentioned b y them I am in th e way to ob ey the orders	Medium	Sprint 2

6. PROJECT PLANNING &SCHEDULING

6.1. Sprint Planning & Estimation

Sprint	**Functional Requirement (Epic)**	**User Story Number**	**User Story / Task Story Points**	**Priority**	**Team Members**
Sprint-1	User Input	USN-1	Test datas are given as input for prediction		

10 High M.Ashmit ha, Priya
L.R.Devi

Sprint-1	USN-2 Model compar es the given data with the Liver disease affect ed data	10	High	J.Thersal, R.Sruthi
Sprint-2	Prediction USN-3 Model predicts the liver disea se using Machin e Learningalgorith m Support Vector Machine(SVM)	10	High	J.Thersal, M.Ashmith a
Sprint-3	Classifier USN-4 Model sends a ll the output to t he classifier and produces the fi nal result.	10	High	R.Sruthi, L.R.Devi Priya

Sprint-4	Announ ce ment	USN-5 Model then displays wheth er the patient is affected by liv er disease or no t	10	High	J.Thersal, R.Sruthi
	Events	USN-6 This model nee ds the capability of displaying accurate result	10	High	M.Ashmit ha, L.R.Devi Priya,

6.2. Sprint Delivery Schedule

Sprint	Total Story Points	Duration	Sprint Start Date	Sprint end Date (Planned)	Story points Completed(as on Planned End Points)	Sprint Release Date (Actual)
Sprint-1	20	6 Days	24 Oct 2022	29 Oct 2022	20	20 Oct 2022
Sprint-2	20	6 Days	31 Oct 2022	05 Nov 2022	20	05 Nov2022

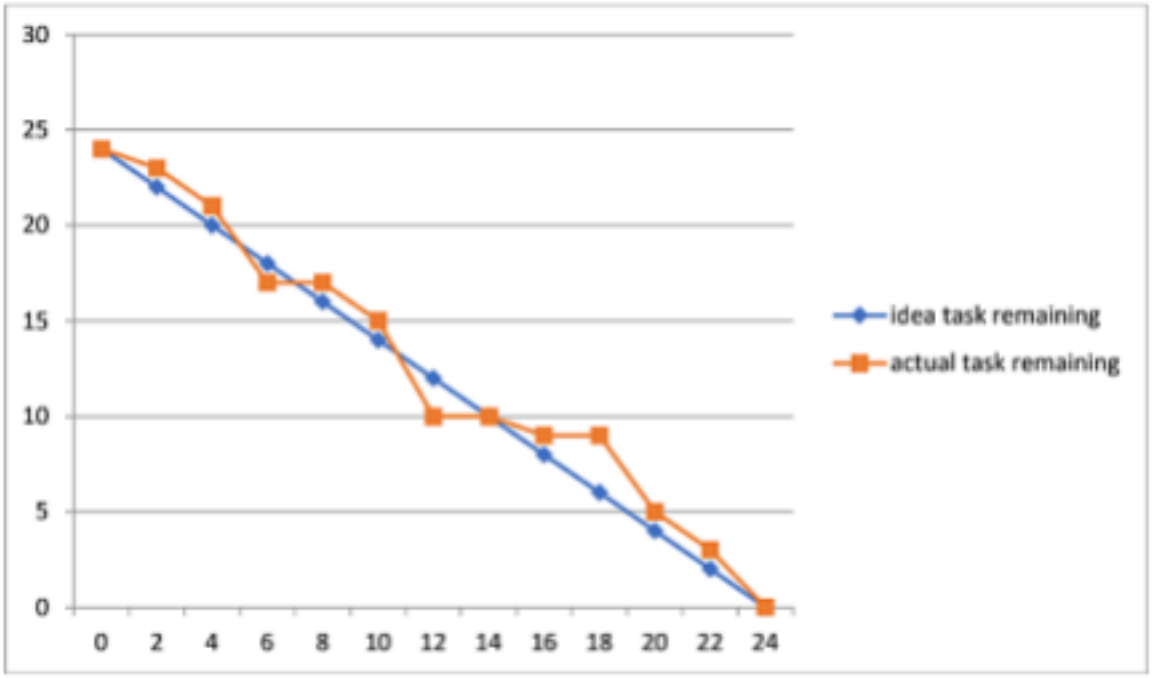
Sprint-3	20	6 Days	07 Nov 2022 12 Nov 2022	20	12 Nov2022
Sprint-4	20	6 Days	14 Nov 2022 19 Nov2022	20	19 Nov2022

Velocity

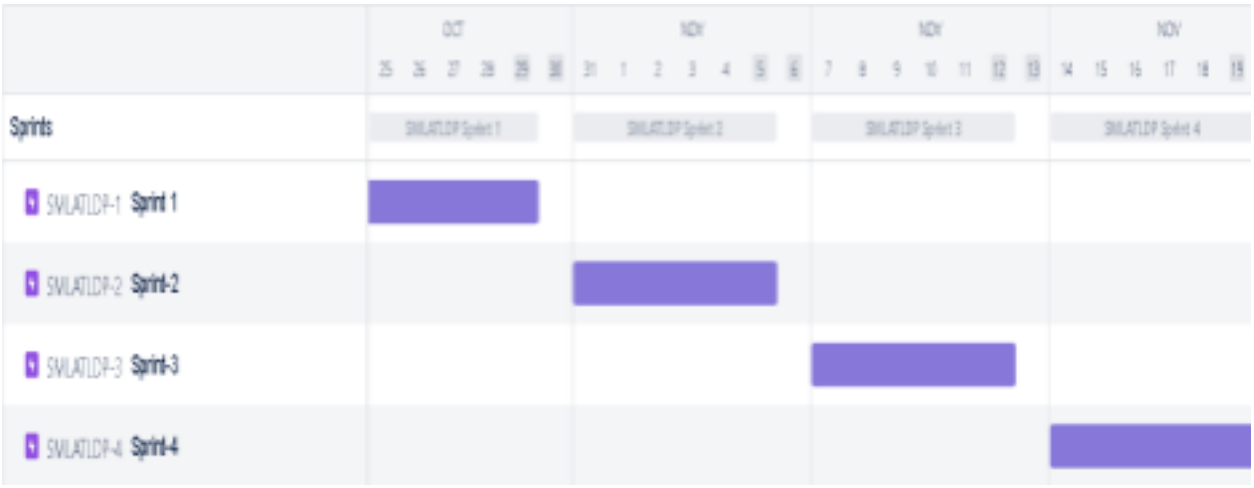
Imagine we have a 10-day spirit duration ,and the velocity of the team is 20(points per spirit). Let's calculate the team average velocity(AV) per iteration unit (story points per day)

AV= Spirit Duration/Velocity =20/10 =2

Burndown Chart



6.3 Reports from JIRA



7.CODING AND SOLUTIONING

7.1 Feature

Login Page: The login page ask the user to enter the data of test result.
Result Page: The result page tells whether the person has liver disease or not.

7.2 Code[s]

home: index

```
<!DOCTYPE html>
<html lang="en">
  <head>
    <meta charset="UTF-8">
    <meta http-equiv="X-UA-Compatible" content="IE=edge">
    <meta name="viewport" content="width=device-width, initial-scale=1.0">
    <title>Document</title>
    <style>
      table
      {
        width:100%;
      }
      td
      {
        height:50px;
```

```
        width:50px;
      }
    </style>
</head>
<body>
    <form method="POST" action="/predict">
      <table>
        <tr>
          <td>
            <label for="age">Age :</label>
          </td>
          <td>
            <input type="number" name="age" step=0.1/>
          </td>
        </tr>
        <tr>
          <td>
            <label for="gender">Gender :</label>
          </td>
          <td>
            <input type="number" name="gender" step=0.1/>
          </td>
        </tr>
        <tr>
          <td>
            <label for="total_bilirubin">Total Bilirubin :</label>
          </td>
          <td>
            <input type="number" name="total_bilirubin" step=0.1/>
          </td>
        </tr>
        <tr>
          <td>
            <label for="alkaline_phosphotase">Alkaline Phosphotase :</label>
          </td>
          <td>
            <input type="number" name="alkaline_phosphotase" step=0.1/>
          </td>
        </tr>
```

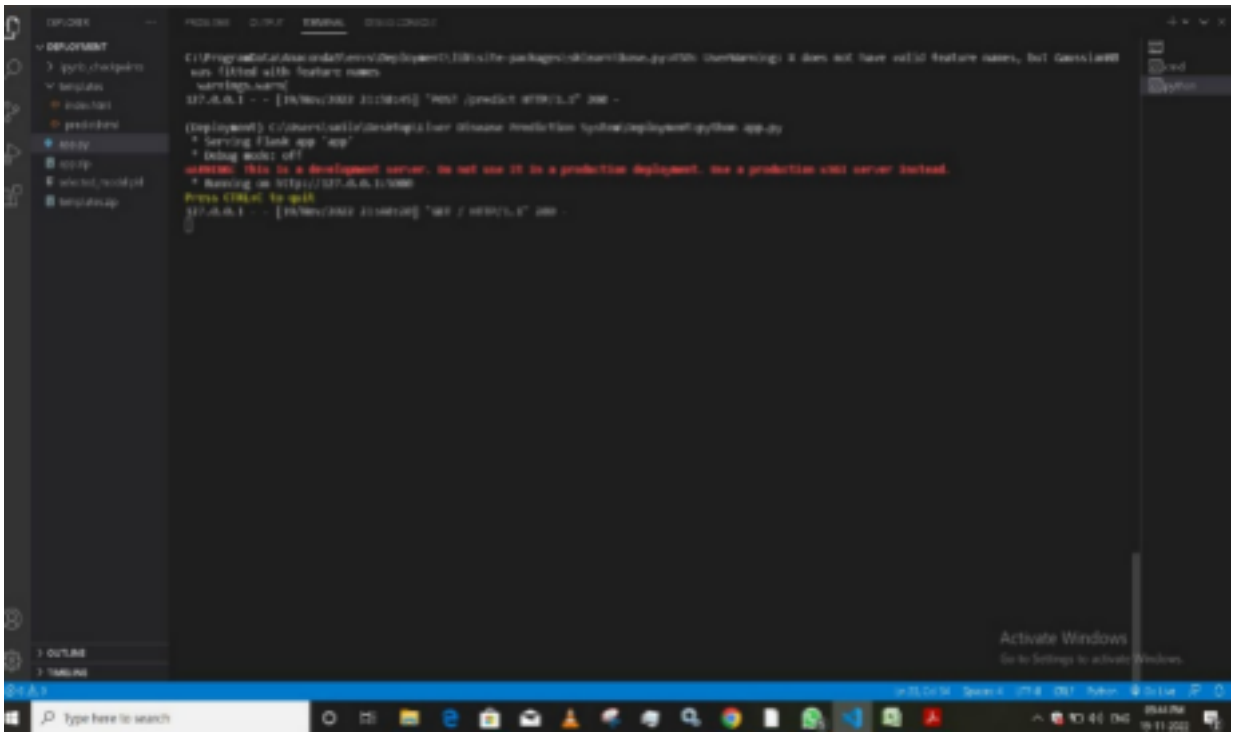
```
            <tr>
              <td>
                <label for="alamine_aminotransferase">Alamine Aminotransferase
:</label>
              </td>
              <td>
                <input type="number" name="alamine_aminotransferase" step=0.1/>
              </td>
            </tr>
            <tr>
              <td>
                <label for="aspartate_aminotransferase">Aspartate Aminotransferase
:</label>
              </td>
              <td>
                <input type="number" name="aspartate_aminotransferase" step=0.1/>
              </td>
            </tr>
            <tr>
              <td>
                <label for="total_protiens">Total Protiens :</label>
              </td>
              <td>
                <input type="number" name="total_protiens" step=0.1/>
              </td>
            </tr>
            <tr>
              <td>
                <label for="albumin">Albumin :</label>
              </td>
              <td>
                <input type="number" name="albumin" step=0.1/>
              </td>
            </tr>
            <tr>
              <td>
                <label for="albumin_and_globulin_ratio">Albumin and Globulin Ratio
:</label>
              </td>
```

```
            <td>
              <input type="number" name="albumin_and_globulin_ratio" step=0.1/>
            </td>
          </tr>
          <tr>
            <td>
            </td>
            <td>
              <button type="submit">Diagonise</button>
            </td>
          </tr>
        </table>
      </form>
  </body>
</html>
```

Predict

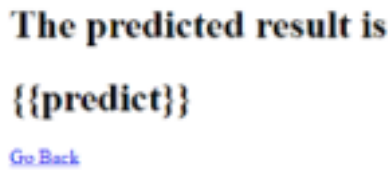
```
        <!DOCTYPE html>
<html lang="en">
<head>
  <meta charset="UTF-8">
  <meta http-equiv="X-UA-Compatible" content="IE=edge">
  <meta name="viewport" content="width=<device-width>, initial-scale=1.0">
  <title>Document</title>
</head>
<body>
  <h1>The predicted result is</h1>
  <h1>{{predict}}</h1>
  <a href="/">Go Back</a>
</body>
</html>
```

8.TESTING



9. RESULTS

In this Project,we found that Liver disease can be Predicted using the test resul t values.



10. ADVANTAGES AND DISADVANTAGES

Advantages:

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 and then you will get the result of prediction.
2. High Accuracy: This system predicts the results with 100% accuracy for the dataset that we have use while creating this application. While the accuracy might be different in some cases, It will still be high enough to be trustworthy at a large scale.
3. Immediate results: The result here are predicted within seconds of entering the details. You dont need to wait for a doctor to come, unlike in traditiona l method.

Disadvantages:

1. Due to any network issue there will be a delay in getting the predicted result
2. It is difficult to implement these techniques in some rural area
3. There is a possibility of entering wrong data so that the predicted result goes wrong

11. CONCLUSION

The mechanisms that are currently used in the prediction of liver disease are prone to have different levels of accuracy and effectiveness. Different diseases demand accuracy of a different set of parameters and might not demand the same set of inferences, throughout more than a single case. In the near future, the study reflects that there was a decent amount of accuracy that was achieved. However, the agenda of our paper is to improvise on those lines and come up with better accuracy standards. The following are some of the clear limitations that have been observed, in order to account for innovation in this paper, having brought about the connotation of improvising on these lines. When it comes to the classification process, it is not necessary that the cohesion that a classifier shares with a particular set of data should stand viable for the rest of the training set. This is to imply that there are some classifiers that don't stand fit to the data set in the context. There are certain methodologies that are incompatible and non-cohesive when

it comes to the collection of real-time data and the implementation procedures of the same. Some of the machine learning approaches that are being considered, do not stand viable for a large volume of data.

12. FUTURE SCOPE

In the future ,we can apply different deep learning and transfer learning algorithms with various feature selection techniques for classifying liver patients and we can use in another set of data and check for the prediction accuracy. And also, we can work on more parameters which help to get better performance .

13. APPENDIX

Source Code

```
        import flask
from flask import request,render_template
from flask_cors import CORS
import joblib
import sklearn

app=flask.Flask(__name__,static_url_path='')
CORS(app)

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

@app.route('/predict',methods=['POST'])
def predictResult():
  a=float(request.form['age'])
  b=float(request.form['gender'])
  c=float(request.form['total_bilirubin'])
  d=float(request.form['alkaline_phosphotase'])
  e=float(request.form['alamine_aminotransferase'
])
  f=float(request.form['aspartate_aminotransferas
e'])   g=float(request.form['total_protiens'])
  h=float(request.form['albumin'])
  i=float(request.form['albumin_and_globulin_ratio'])
  x=[[a,b,c,d,e,f,g,h,i]]
  model=joblib.load('selected_model.pkl')
  result=model.predict(x)[0]
  if(result==2):
    res="Liver Disease Predicted"
  else:
    res="No Liver Disease Predicted"
  return render_template('predict.html',predict=res)

if __name__=='__main__':
  app.run()
```

Github:

https://github.com/IBM-EPBL/IBM-Project-40021-1660614662