EARLY DETECTION OF CHRONIC KIDNEY DISEASE USING MACHINE LEARNING

PROJECT REPORT

TEAM ID: PNT2022TMID34602

TEAM LEAD: ASHA LAXMI M A

TEAM MEMBERS: BAMIYA J RENISH

DARINCE C

HESSIYA K K

CONTENT

1 INTRODUCTION

1.1 Project Overview:

Chronic kidney Disease (CKD) means your kidneys are dam-aged and not filtering your blood the way it should. The primary role of kidneys is to filter extra water and waste from your blood to produce urine and if the person has suffered from CKD, it means that wastes are collected in the body. This disease is chronic because of the damage gradually over a long period. It is flattering a common disease world wide. Due to CKD may have some health troubles. There are many causes for CKD like diabetes, high blood pressure, heart disease

1.2 Purpose:

Chronic Kidney Disease (CKD) or chronic renal disease has become a major issue with a steady growth rate. A person can only survive without kidneys for an average time of 18 days, which makes a huge demand for a kidney transplant and Dialysis. It is important to have effective methods for early prediction of CKD. Machine learning methods are effective in CKD prediction. This work proposes a workflow to predict CKD status based on clinical data, incorporating data prepossessing, a missing value handling method with collaborative filtering and attributes selection. Out of the 11 machine learning methods considered, the extra tree classifier and random forest classifier are shown to result in the highest accuracy and minimal bias to the attributes. The research also considers the practical aspects of data collection and highlights the importance of incorporating domain knowledge when using machine learning for CKD status prediction.

2 LITERATURE SURVEY

2.1 Existing problem:

While predicting the CKD in early stages using multi-layer perception while including preprocessing of dataset with neural networks to fill the missing values. The work flow includes discarding the outliers ,selecting the optimal seven attributes with statistical analysis, and discarding the attributes which have inter co-relation by principle component analysis. .In the mentioned work,the missing value filling algorithm has a significant impact on the accuracy of the trained

models.However ,because of using Neural Network for 20 attributes only with 260 fully completed data instances,the accuracy of missing value prediction is sightly reduced.

2.2 Reference

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9321365/

2.3 Problem statement definition:

Chronic kidney disease (CKD) is a worldwide public health problem, with adverse outcomes of kidney failure, cardiovascular disease (CVD), and premature death. A simple definition and classification of kidney disease is necessary for international development and implementation of clinical practice guidelines. CKD is defined as kidney damage or glomerular filtration rate (GFR) <60 mL/min/1.73 m (2) for 3 months or more, irrespective of cause. Kidney damage in many kidney diseases can be ascertained by the presence of albuminuria, defined as albumin-to-creatinine ratio >30 mg/g in two of three spot urine specimens.

1 IDEATION AND PROPOSED SOLUTION

3.1 Empathy map canvas

3.2 Ideation & brainstorming:

Ideation is the process where you generate ideas and solutions through sessions such as Sketching, Prototyping, Brainstorming, Brainwriting, Worst Possible Idea, and a wealth of other ideation techniques. Ideation is also the third stage in the Design Thinking process.

Brainstorming is a method of generating ideas and sharing knowledge to solve a particular commercial or technical problem, in which participants are encouraged to think without interruption. Brainstorming is a group activity where each participant shares their ideas as soon as they come to mind.

3.3 Proposed Solution:

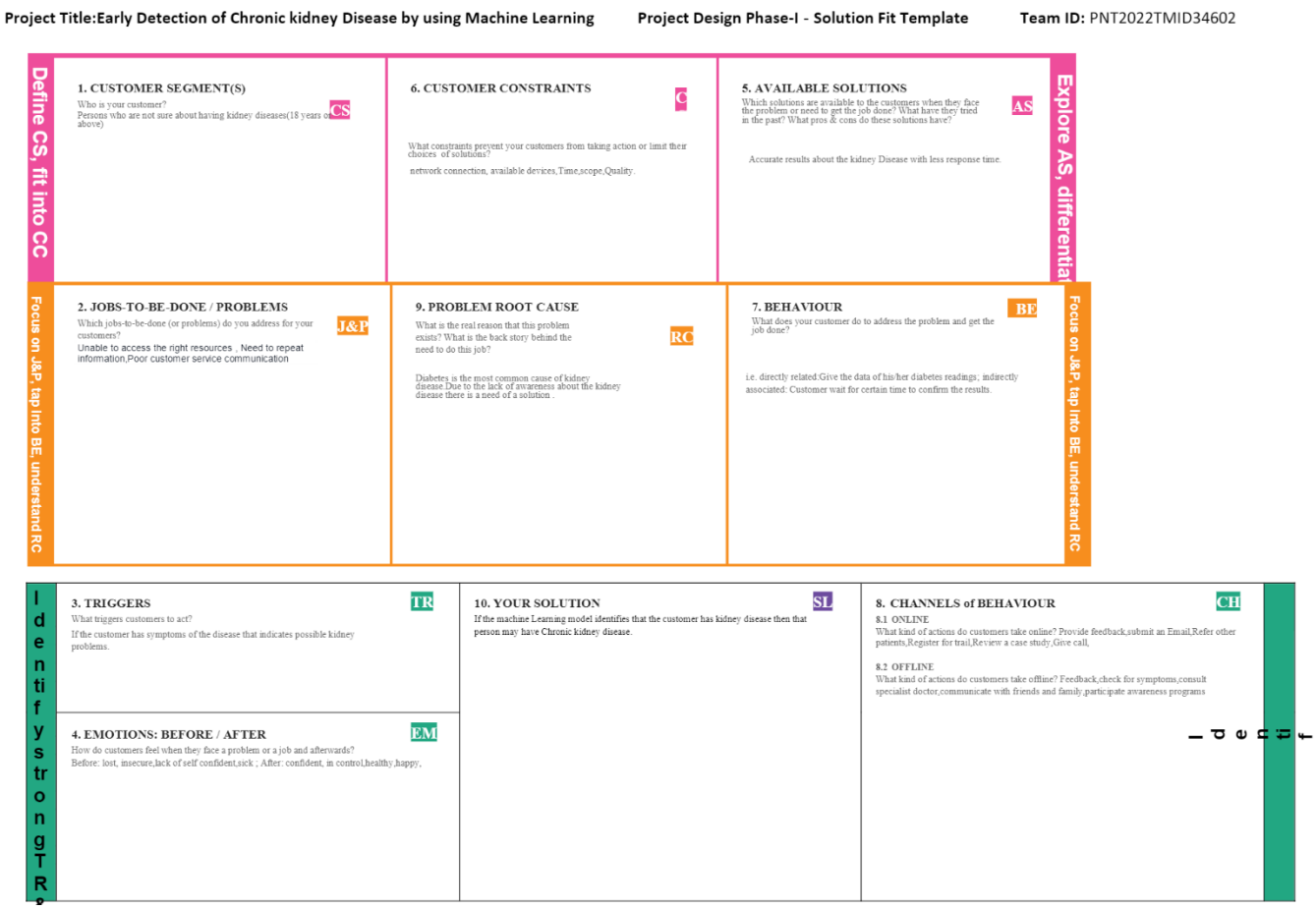
The Proposed System is an mechanism for predicting Chronic Kidney Disease using different Classification Techniques. CKD can be Classified according to its severeness using Machine learning process. Aim of this project is to predict CKD using different Machine learning algorithms, medical test records of CKD patient can be utilized to recommend diet plan by using different classification algorithm. System uses old data from “UCI Repository” and uses tools such as “Visual Studio” and “SQL Server” to

develop application. System is an real time application useful for doctors to identify CKD and related stages and recommending the suitable diet for the patients

3.4 Problem Solution Fit:

We have developed a model to predict CKD disease in patients. The performance of the model was tested on both all attributes and selected features. Among feature selection methods there were Wrapper, Filter and Embedded allowing to select vital features. Classifier algorithms performance was tested on the selected features. IBM SPSS tool is used for preparing the model. The machine learning classifiers such as artificial neural network (ANN), C5.0, logistic regression, linear support vector machine (LSVM), K- nearest neighbors (KNN) and random tree were used for training the model.

.



4 REQUIREMENT ANALYSIS

4.1 Functional requirement:

Patient data sets and parameters: In the first step of prediction process where we collect medical data. datasets were used for processing. Training data-sets will contains patient details and also parameters that are required for prediction which is shown in Table 1 and Table 2, where table 1 describes the attributes and their description used in the CKD prediction data sets. The above table 2 describes the attributes and their measurements used in the CKD prediction data sets.

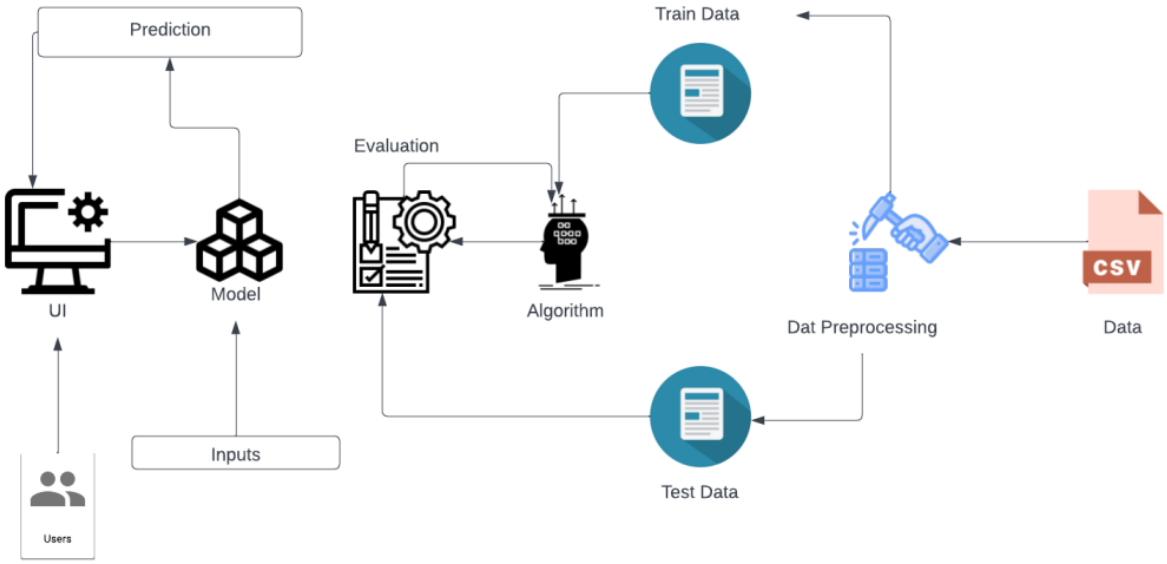
Extract and Segment Data (Data Pre-processing): The medical data is analysed and only relevant dataset are extracted. The data required for processing is extracted and segmented according to the requirement. This is done because entire training data not required for processing and if we input all data, it requires too much of time for processing, so data processing is done.

4.2 Non-Functional requirement:

1. Performance and scalability
2. Portability and compatibility.
3. Reliability, maintainability, availability.
4. Usability.

5 PROJECT DESIGN

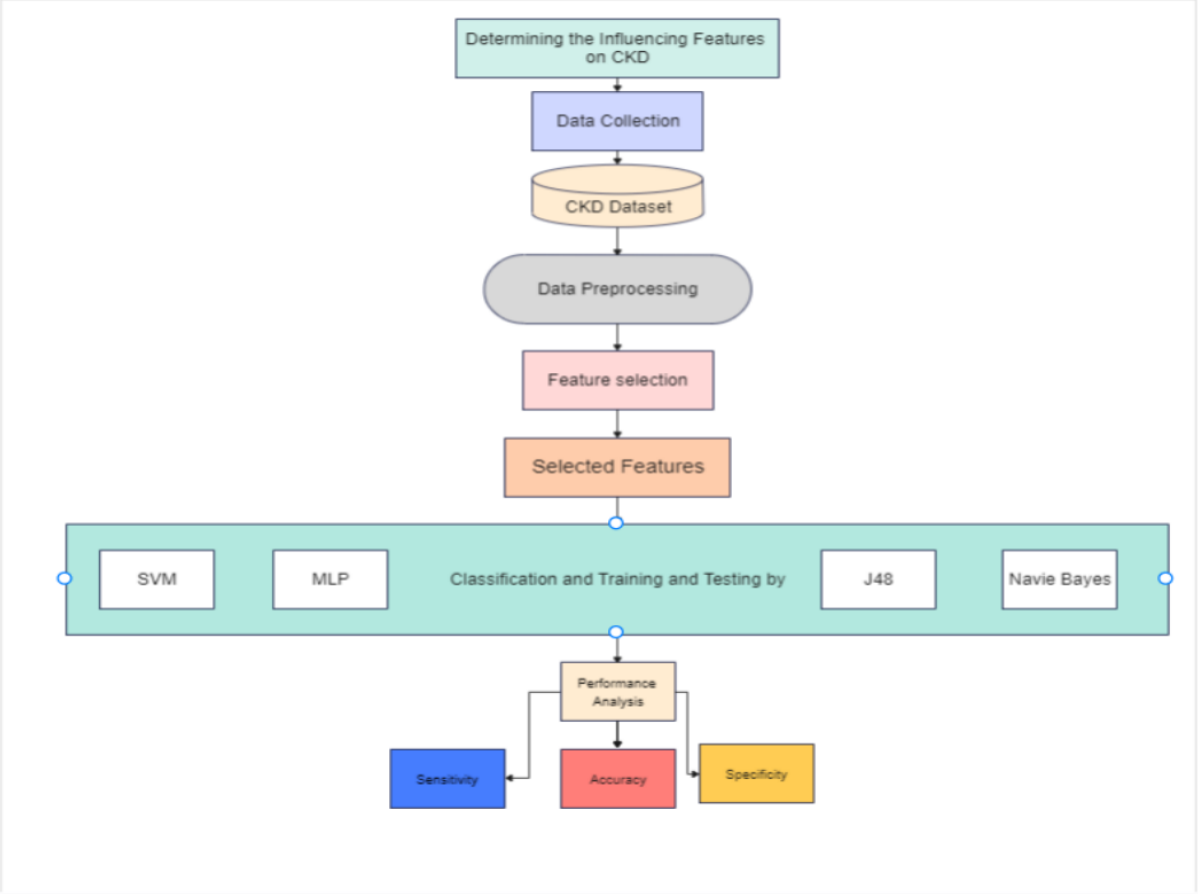
5.1 Data Flow Diagram:



5.2 Solution & Technical Architecture:

Solution architecture is a complex process – with many sub-processes – that bridges the gap between business problems and technology solutions. Its goals are to:

- Find the best tech solution to solve existing business problems.
- Describe the structure, characteristics, behaviour, and other aspects of the software to project stakeholders.
- Define features, development phases, and solution requirements.
- Provide specifications according to which the solution is defined, managed, and delivered.



5.3 User Stories:

- As a user, I can read the instructions and data provided in the dashboard and come to a clear view about chronic kidney disease.
- As a user, I can read the instructions and data provided in the dashboard and come to a clear view about chronic kidney disease.
- As a user, I can be strong about my prediction by the model.
- As a user, I can expect exact level of accuracy if I have the disease or not.
- As a user, I need to enter the data for prediction.

6 PROJECT PLANNING AND SCHEDULING

6.1 Sprint planning and Estimation:

- Sprint 1- Data collection

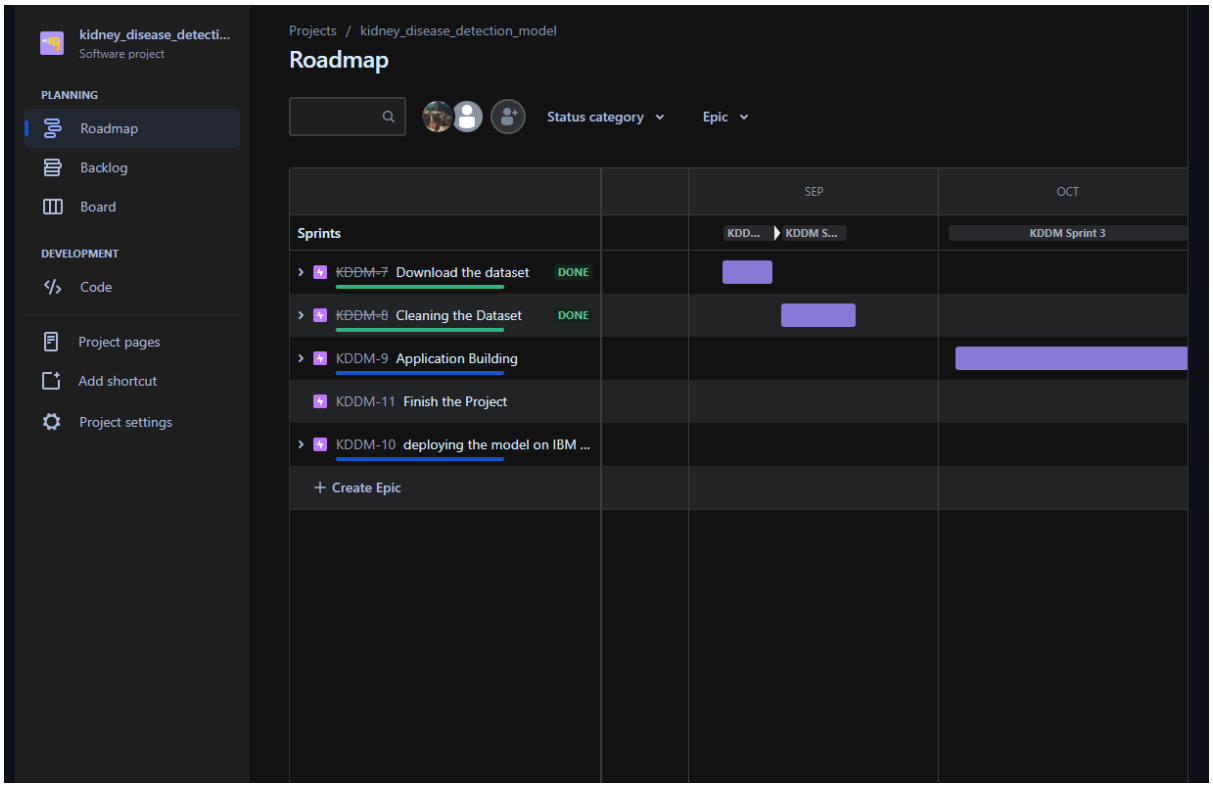
- Estimation – 05/09/2022-12/09/2022

- Sprint 2- Cleaning the dataset
  - Estimation -12/09/2022-20/09/2022

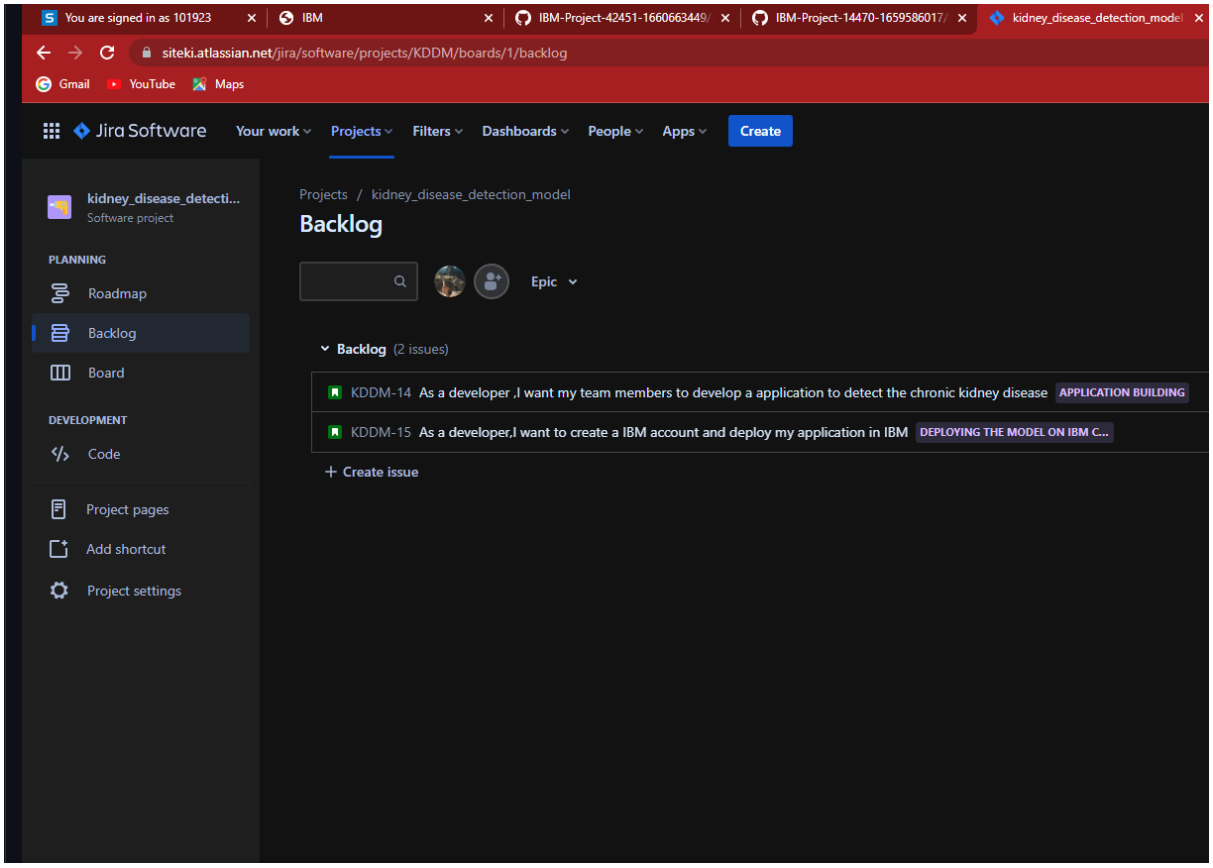
- Sprint 3- Application building
  - Estimation-02/10/2022-30/10/2022

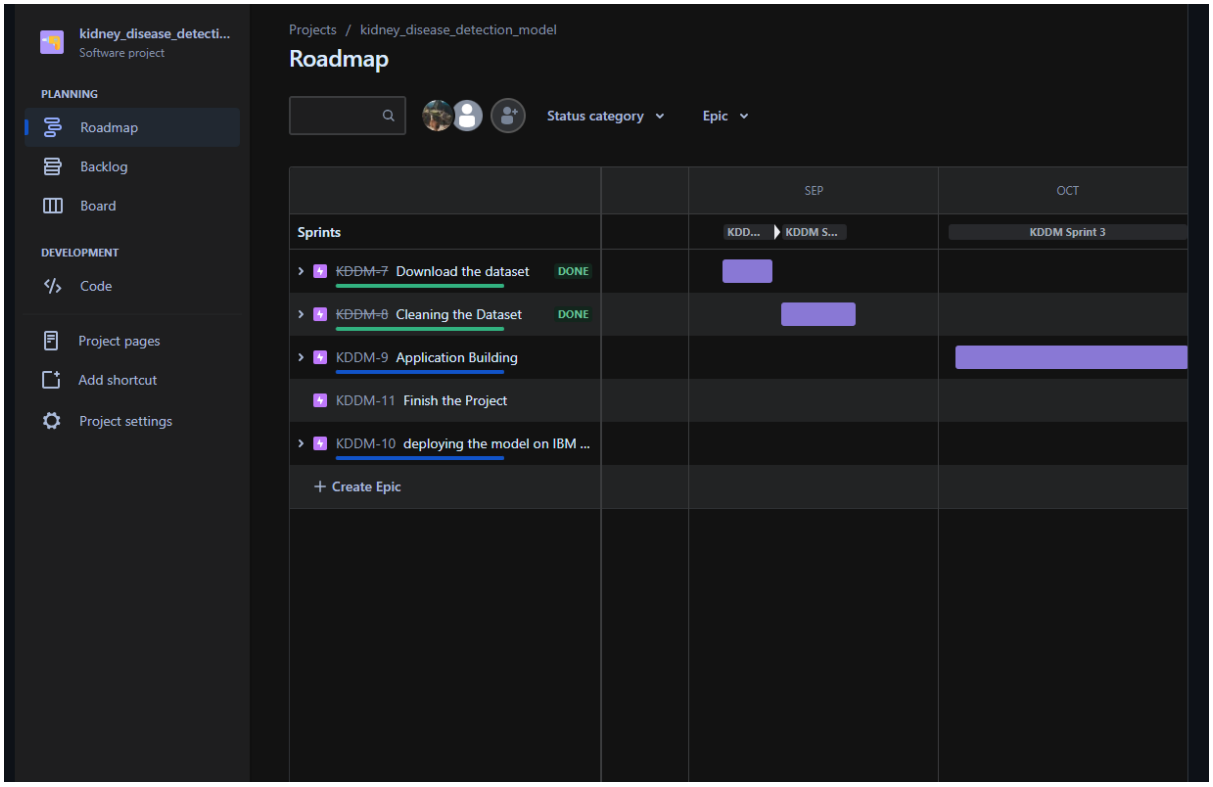
- Sprint 4- Deploying the model on IBM cloud
  - Estimation - 08/11/2022-15/11/2022

6.2 Sprint delivery schedule:



6.3 Report From JIRA:


You are signed in as 101923
IBM
IBM-Project-42451-1660663449/
IBM-Project-14470-1659586017/
kidney_disease_detection_model
siteki.atlassian.net/jira/software/projects/KDDM/boards/1/backlog
Gmail
YouTube
Maps
Jira Software
Your work
Projects
Filters
Dashboards
People
Apps
Create
kidney_disease_detecti...
Software project
PLANNING
Roadmap
Backlog
Board
DEVELOPMENT
Code
Project pages
Add shortcut
Project settings
Projects / kidney_disease_detection_model
Backlog
Epic
Backlog (2 issues)
KDDM-14 As a developer ,I want my team members to develop a application to detect the chronic kidney disease
APPLICATION BUILDING
KDDM-15 As a developer,I want to create a IBM account and deploy my application in IBM
DEPLOYING THE MODEL ON IBM C...
Create issue



7 CODING AND SOLUTIONING

7.1 Feature 1:

The main feature of the chronic kidney disease prediction model is that it efficiently predicts the person suffering from chronic kidney disease and also to identify the healthy perosn by screening the diabetes mellitus levels, albumin levels etc.

Code:

```
{% if pred == 1 %}
    <div class="jumbotron ">
      <h1 class="display-4">You have a Kidney Disease !</h1>
      <p class=”lead”>Please Consult the Doctor Immediately.It was too risky without Consultation.. Make sure of health in your diet</p>
      <p>Proper consultation needed</p>
```

7.2 Feature 2:

Other features of the chronic kidney disease prediction model is that the model checks for disease ,also the end user can get the opportunity to be aware about kidney disease and related symptoms.

Code:

```
<section class="jumbotron p-3 p-md-5 text-white rounded bg-dark text-
center">
        <div class="container">
          <h1 class="jumbotron-heading">Chronic Kidney Disease
Prediction</h1>
          <p class="lead">Chronic kidney disease (CKD) is one of the most
            critical health problems due to its increasing prevalence. In this
            paper, we aim to test the ability of machine learning algorithms
            for the prediction of chronic kidney disease using the smallest
            subset of features</p>
          <p>
            <a href="https://www.mayoclinic.org/diseases-conditions/chronic-
kidney-disease/symptoms-causes/syc-20354521"
                class="btn btn-primary my-2">Read More about the
Disease</a>
          </p>
        </div>
      </section>
```

7.3 Database Scheme:

There is no database schema used in this web application. But flask has the default database of sqlite database.

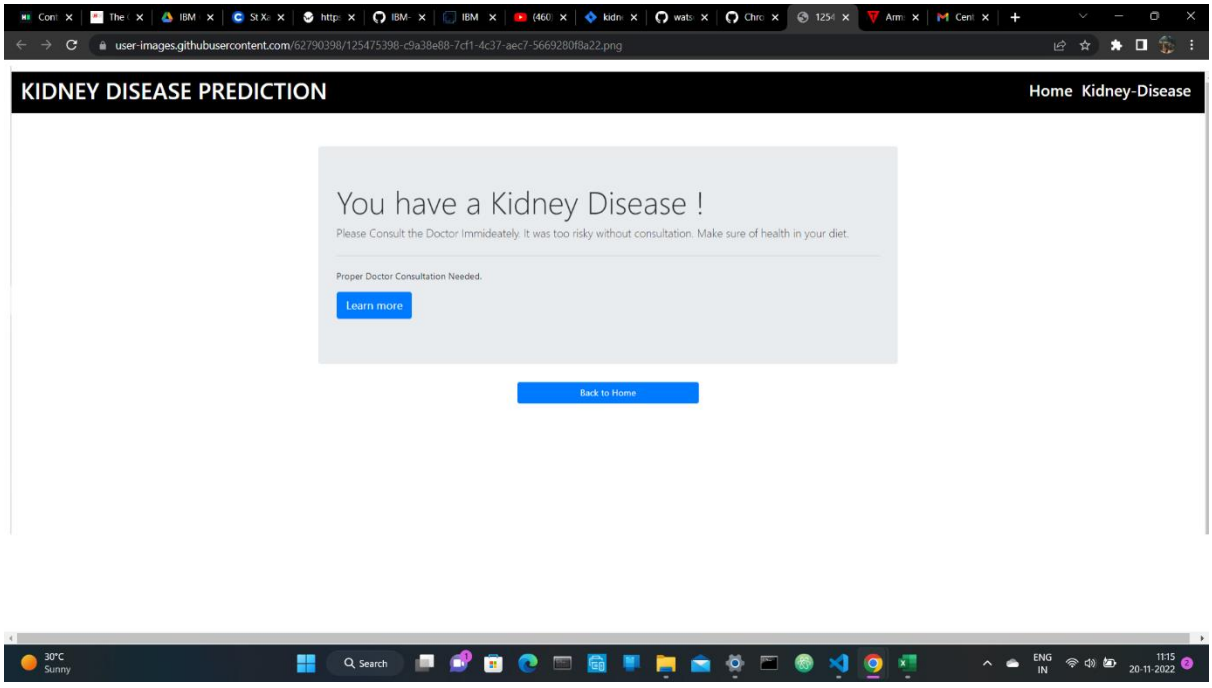
8 TESTING

8.1 Test Case:

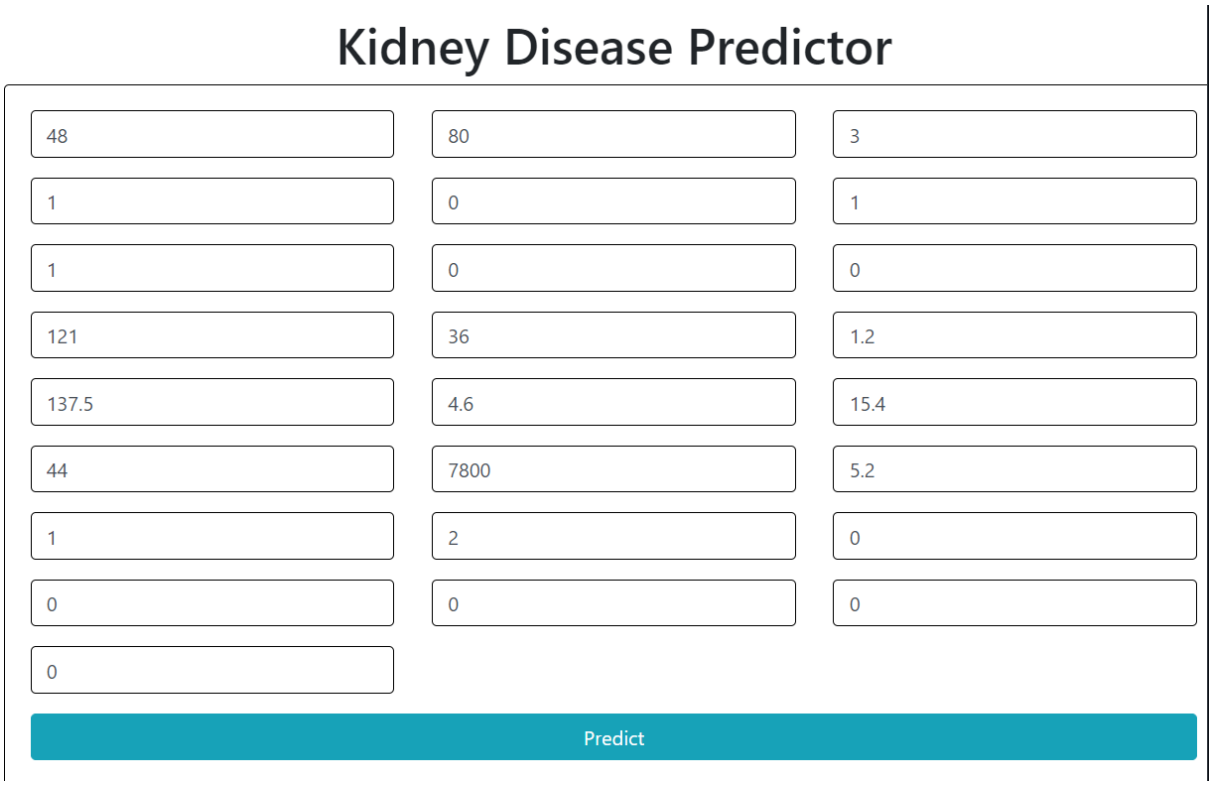
Kidney Disease Predictor

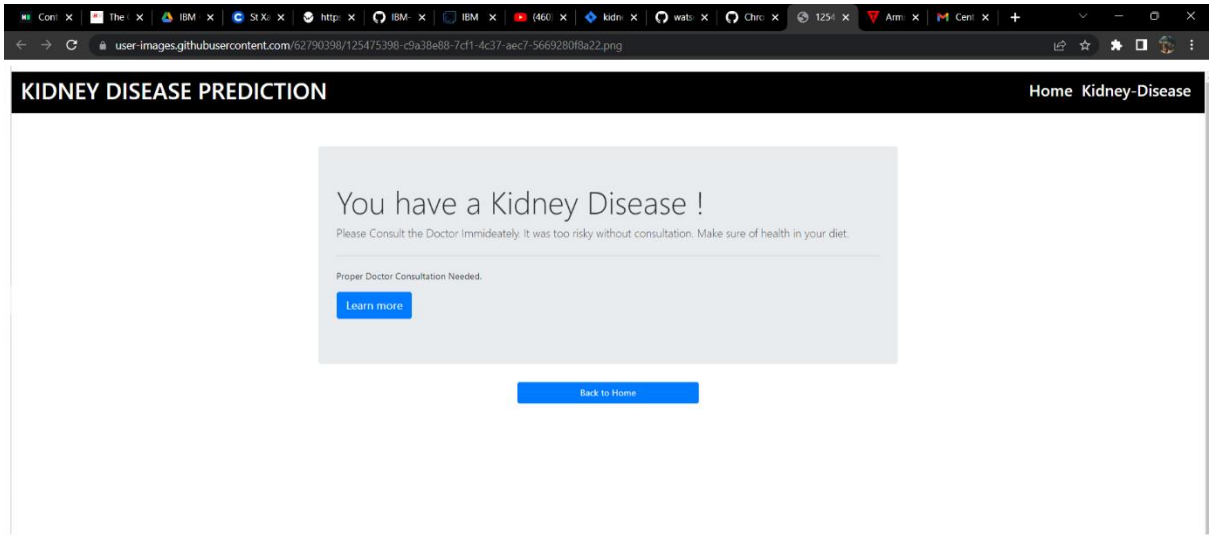
48	80	3
1	0	1
1	0	0
121	36	1.2
137.5	4.6	15.4
44	7800	5.2
1	2	0
0	0	0
0		

Predict



8.2User Acceptance Testing :





9 RESULT

9.1 Performance metrices:

Performance metrics are a part of every machine learning pipeline. They tell us if we're making progress and put a number on it .All machine learning models, whether it's linear regression, or SOTA technique like BERT,need a metric to judge performance.

Confusion Matrix

Confusion Matrix is a tabular visualization of the **ground-truth labels versus model predictions**. Each row of the confusion matrix represents the instances in a predicted class and each column represents the instances in an actual class. Confusion Matrix is not exactly a performance metric but sort of a basis on which other metrics evaluate the results.

In order to understand the confusion matrix, we need to set some value for the null hypothesis as an assumption. For example, from our Breast Cancer data, let's assume our **Null Hypothesis H^0** be "*The individual has cancer*".

In our program the confusion matrix values has been as follows,

```
confusion_matrix(y_test, model.predict(X_test))
```

```
array([[23, 0],
       [ 0, 9]], dtype=int64)
```

10 ADVANTAGES AND DISADVANTAGES

Advantages:

Until now, in majority of cases full features have been taken into consideration. In this research, feature optimization was carried out, wherein three different feature selection algorithms were applied to find the algorithm most beneficial to extract the important feature for the prediction of Chronic Kidney Disease

Disadvantages:

As many datasets have imbalanced class, class balancing is needed for increasing the performance of classifier model.

11 CONCLUSION

This article objects to predict Chronic Kidney Disease based on full features and important features of CKD dataset, For feature selection three different technique have been applied: correlation-based feature selection, Wrapper method and LASSO regression.

12 FUTURE SCOPE

The full features of the dataset were used and the result was tested on all seven machine learning classification algorithms with 50% of training data and 50%of testing data. The comparison matrix was created for all algorithms.

With the resultant matrix, three graphs were also created for checking the variation of various classifiers. The first graph provides a comparison of all classifier's accuracy, precision and recall.

The second one contains the variation of AUC and the third one includes the variation of F-measure. The comparison of all classifiers showed that the C5.0 algorithm achieved the highest accuracy, i.e. 96.10%.

13 APPENDIX

Source code:

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

```
from tensorflow.keras.models import load_model



app = Flask(__name__)


def predict(values, dic):
    if len(values) == 18:
        model = pickle.load(open('models/kidney.pkl', 'rb'))
        values = np.asarray(values)
        return model.predict(values.reshape(1, -1))[0]



@app.route("/")
def home():
    return render_template('home.html')
@app.route("/kidney", methods=['GET', 'POST'])
def kidneyPage():
    return render_template('kidney.html')

@app.route("/predict", methods=['POST', 'GET'])
def predictPage():
    try:
        if request.method == 'POST':
            to_predict_dict = request.form.to_dict()
```

```
            to_predict_list = list(map(float, list(to_predict_dict.values())))
            pred = predict(to_predict_list, to_predict_dict)
        except:
            message = "Please enter valid Data"
            return render_template("home.html", message=message)

        return render_template('predict.html', pred=pred)


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

GITHUB: https://github.com/IBM-EPBL/IBM-Project-42451-1660663449