EARLY DETECTION OF CHRONIC KIDNEY DISEASE USING MACHINE LEARNING

PROJECT BASED LEARNING (NALAIYA THIRAN)
On
PROFESSIONAL READINESS FOR INNOVATIONS, EMPLOYABILITY AND ENTREPRENEURSHIP

Submitted by

TEAM ID: PNT2022TMID38667

TEAM MEMBERS:

SAHANA R	420419205011 (TL)
KANIMOZHI D	420419205006 (TM1)
KEERTHANA V	420419205007 (TM2)
TAMILARASI S	420419205017 (TM3)
BHAVANI G	420419205301 (TM4)

in partial fulfilment for the award of the degree

of

BACHELOR OF TECHNOLOGY

IN

INFORMATION TECHNOLOGY

ADHIPARASAKTHI ENGINEERING COLLEGE,

MELMARUATHUR

ANNA UNIVERSITY: CHENNAI – 600 025

NOVEMBER - 2022

ANNA UNIVERSITY: CHENNAI 600 025

BONAFIDE CERTIFICATE

Certified that report titled **“EARLY DETECTION OF CHRONIC KIDNEY DISEASE USING MACHINE LEARNING”** is the bonafide work of **SAHANA R(420419205011), KANIMOZHI D(420419205006), KEERTHANA V(420419205007),TAMILARASI S(420419205017) and BHAVANI G(420419205301)** who carried out the work under my supervision.

SIGNATURE	**SIGNATURE**	**SIGNATURE**
Dr.A.Bhuvaneswari M.E.,Ph.D.,	**Dr.N.Elamathi M.E.,Ph.D.,**	**Mr.M.Ezhilvendan M.E.,**
HOD and SPOC	**Faculty Mentor**	**Internal Evaluator**
Professor,	**Assistant Professor,**	**Assistant Professor,**
Department of	Department of	Department of
Information Technology	Information Technology,	Information Technology,
Adhiparasakthi Engineering College,	Adhiparasakthi Engineering College,	Adhiparasakthi Engineering College,
Melmaruvathur - 603319	Melmaruvathur - 603319	Melmaruvathur - 603319

CERTIFICATION OF EVALUATION

College Code/Name	**:** 4204 / Adhiparasakthi Engineering College
Branch/ Semester	**:** Information Technology /07
Team ID	**:** PNT2022TMID38667

S.No	Name of the student and Register number	Title of the project	Name of the Faculty Mentor with designation
1 2 3 4 5	**SAHANA R** (420419205011) **KANIMOZHI D** (420419205006) **KEERTHANA V** (420419205007) **TAMILARASI S** (420419205017) **BHAVANI G** (420419205301)	Early detection of chronic kidney disease using machine learning.	Dr.N.Elamathi M.E.,Ph.D., Assistant Professor, Department of IT, Adhiparasakthi Engineering College, Melmaruvathur

The report of the project works submitted by the above students in the partial fulfillment for the award of Bachelor of Technology degree in Information Technology of Anna University were evaluated and confirmed to be reports of work done by the above students and then evaluated

Submitted for the Project Work and Viva -voce examination held on……………..

INTERNAL EXAMINER **EXTERNAL EXAMINER**

ACKNOWLEDGEMENT

It is indeed a great pleasure and proud privilege to acknowledgement the help and support we received from the positive minds around us in making this Endeavour a successful one the spiritual blessings of His Holiness **ARULTHIRU AMMA** and the divine guidance of **THIRUMATHI AMMA** have undoubtedly taken us to the path of victory in completing this project.

The infrastructural support with all kinds of lab facilities have been a motivating factor in this completion of project work, all because of our **CORRESPONDENT SAKTHI THIRU Dr. G. B. SENTHILKUMAR** with great pleasure we take this opportunity to thank him.

From the academic side the constant support from our honourable **PRINCIPAL Dr. J. RAJA, Ph.D.,** has encouraged us to work hard and attain this goal of completing the project. We sincerely thank our motivating and respected **HEAD OF THE DEPARTMENT and SPOC**

Dr. A. BHUVANESWARI M.E., Ph.D., who have given us both moral and technical support adding experience to the job we have undertaken.

We are thanking our respected Dr.N.Elamathi M.E.,Ph.D., **FACULTY MENTOR, Mr.M.Ezhilvendan M.E., INTERNAL EVALUATOR** and Assistant Professor who helped us in crossing obstacles in the path to our glory. We also thank other Staff members, Non- teaching members of Main Block Computer Lab. Parents and Friends who have given their constant support and motivation in all our endeavours.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

Chronic kidney disease prediction is one of the most important issues in health care-analytics. The most interesting and challenging tasks in day-to-day lives as one third of adult population is affected by chronic kidney disease (CKD), and millions die each year because they do not have access to affordable treatment. Chronic Kidney Disease can be cured, if treated in the early stages. The main aim of the project is to predict whether the patient have chronic kidney disease or not in a painless, accurate and faster way based on certain diagnostic measurement like Blood Pressure (BP), Albumin (Al) etc., and then appropriate treatment can be given based on the details provided by the model.

1.1 Project Overview:

This Project aims at creating a model for early detection of Chronic Kidney Disease using Machine Learning technology. The model output is integrated with Flask framework. The front end developed in html is used to receive user input on various parameters needed to decide on the early detection of kidney disease. The same model is deployed into IBM cloud using API keys and scoring endpoints.

1.2 Purpose:

- The goals of early detection are to prevent the progression of chronic kidney disease and its associated complications, with subsequent improvements in patient outcomes and reductions in the impact of chronic kidney disease on healthcare resources.

- The purpose of the project is to alert doctors for an early detection of kidney disease and hence ensure speedy recovery or prevention of kidney disease.

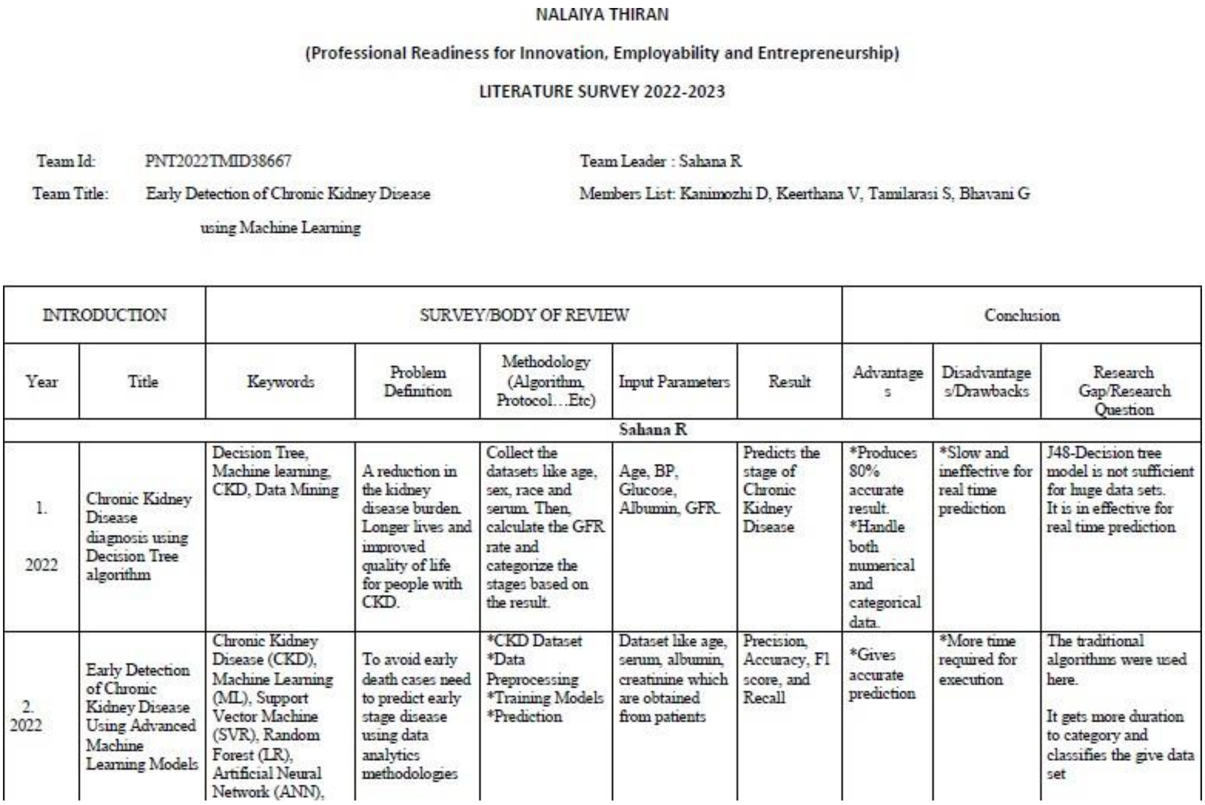
2. LITERATURE SURVEY

Initially, we have done literature survey of various IEEE papers and research publications to arrive at the idea of the project development. It is given below:

2.1 Existing Solution:

The current existing solutions offer a methodology for predicting CKD status using clinical data, which incorporates data pre-processing, a technique for managing missing values, data aggregation, and feature extraction. A number of physiological variables, as well as ML techniques such as logistic regression (LR), decision tree (DT) classification, and -nearest neighbour (KNN), were used in this work to train three distinct models for reliable prediction.

2.2 Reference:



NALAIYA THIRAN

(Professional Readiness for Innovation, Employability and Entrepreneurship)

LITERATURE SURVEY 2022-2023

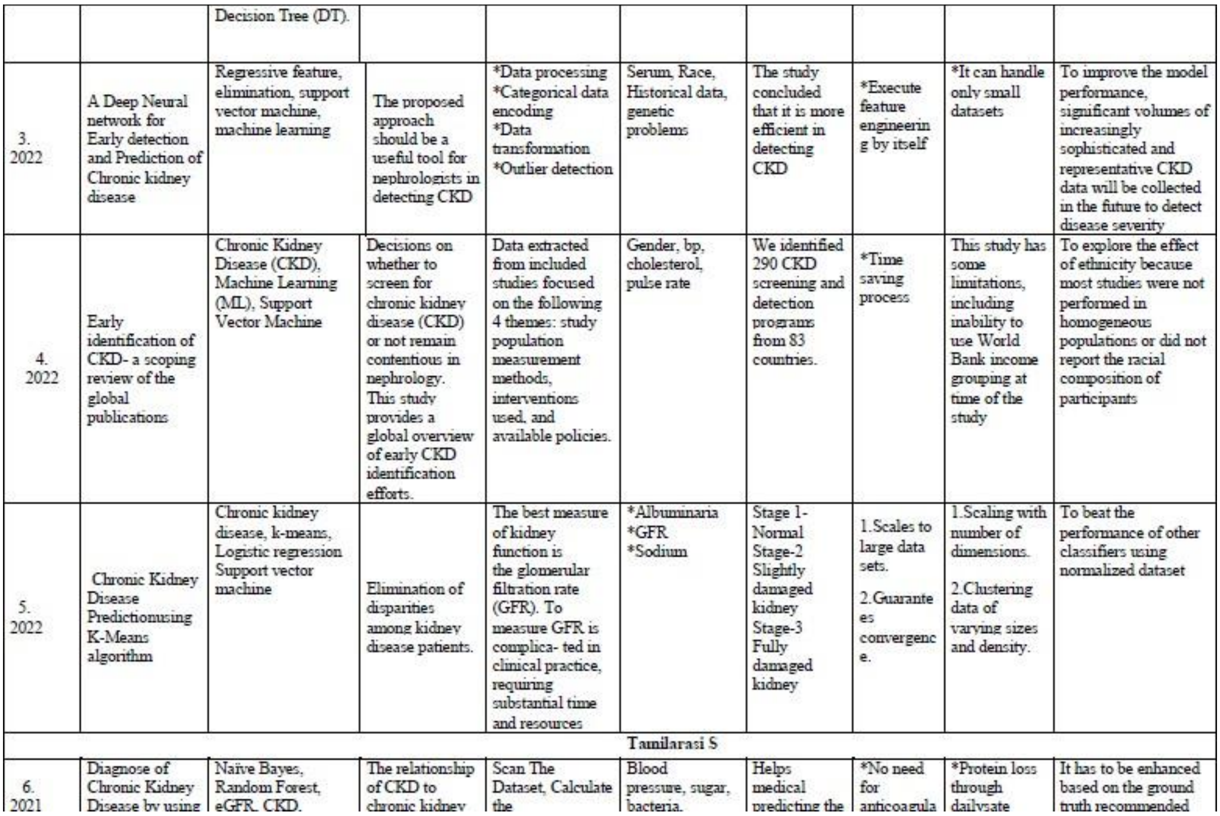
Team Id: PNT2022TMID38667

Team Title: Early Detection of Chronic Kidney Disease using Machine Learning

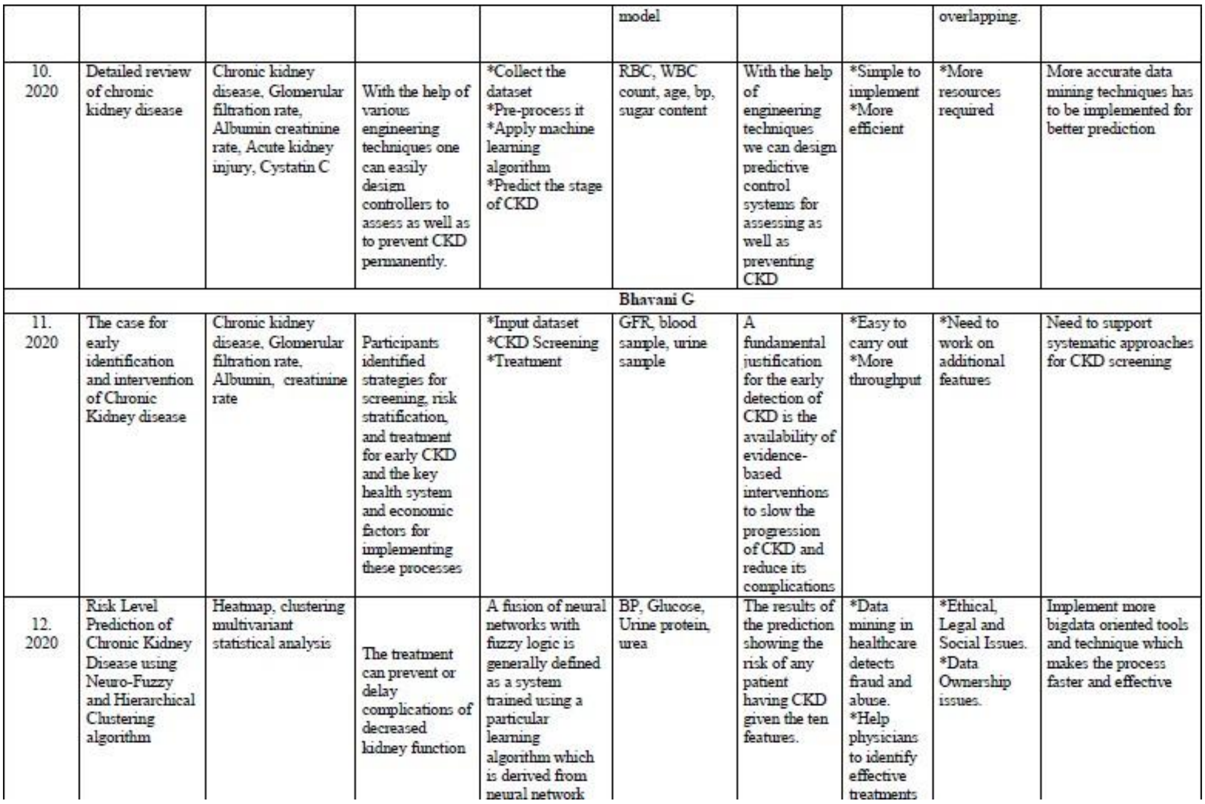
Team Leader : Sahana R

Members List: Kanimozhi D, Keerthana V, Tamilarasi S, Bhavani G

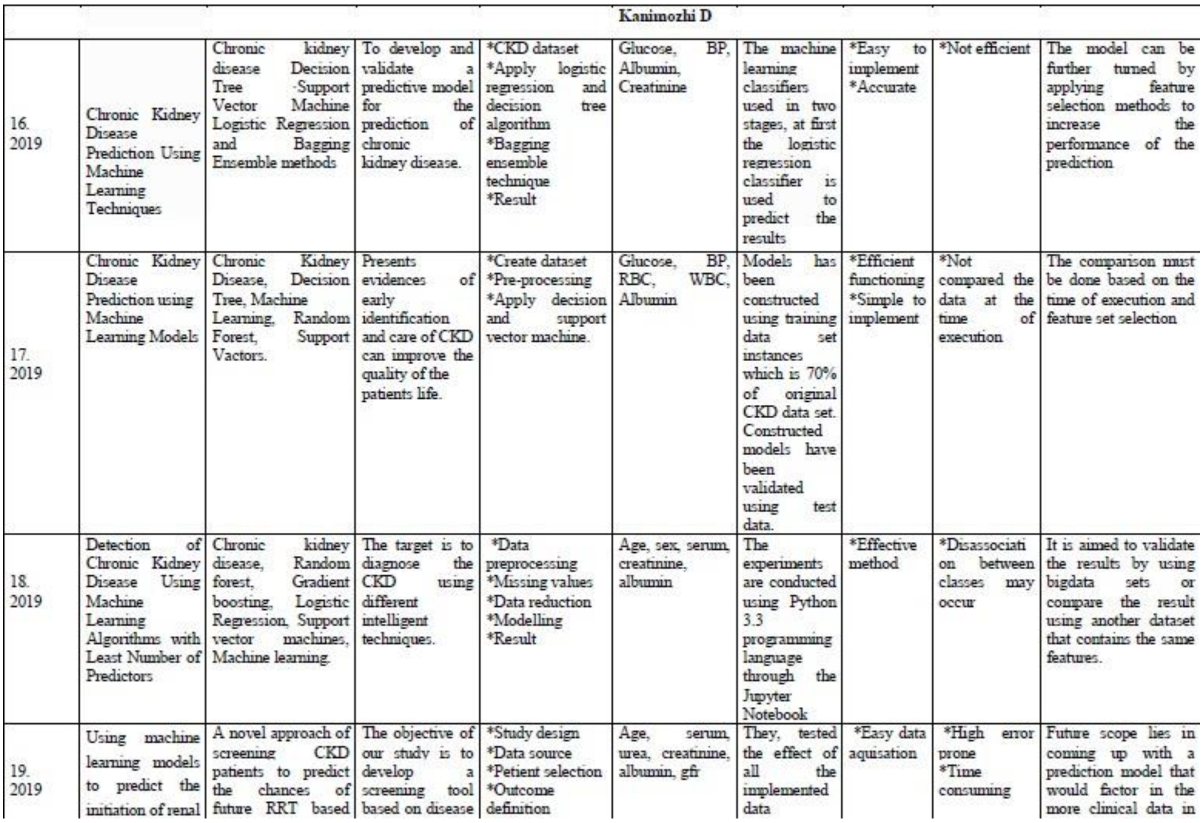
INTRODUCTION		SURVEY/BODY OF REVIEW					Conclusion		
Year	Title	Keywords	Problem Definition	Methodology (Algorithm, Protocol…Etc)	Input Parameters	Result	Advantages	Disadvantages/Drawbacks	Research Gap/Research Question
					Sahana R				
1. 2022	Chronic Kidney Disease diagnosis using Decision Tree algorithm	Decision Tree, Machine learning, CKD, Data Mining	A reduction in the kidney disease burden. Longer lives and improved quality of life for people with CKD.	Collect the datasets like age, sex, race and serum. Then, calculate the GFR rate and categorize the stages based on the result.	Age, BP, Glucose, Albumin, GFR.	Predicts the stage of Chronic Kidney Disease	*Produces 80% accurate result. *Handle both numerical and categorical data.	*Slow and ineffective for real time prediction	J48-Decision tree model is not sufficient for huge data sets. It is in effective for real time prediction
2. 2022	Early Detection of Chronic Kidney Disease Using Advanced Machine Learning Models	Chronic Kidney Disease (CKD), Machine Learning (ML), Support Vector Machine (SVR), Random Forest (LR), Artificial Neural Network (ANN),	To avoid early death cases need to predict early stage disease using data analytics methodologies	*CKD Dataset *Data Preprocessing *Training Models *Prediction	Dataset like age, serum, albumin, creatinine which are obtained from patients	Precision, Accuracy, F1 score, and Recall	*Gives accurate prediction	*More time required for execution	The traditional algorithms were used here. It gets more duration to category and classifies the give data set

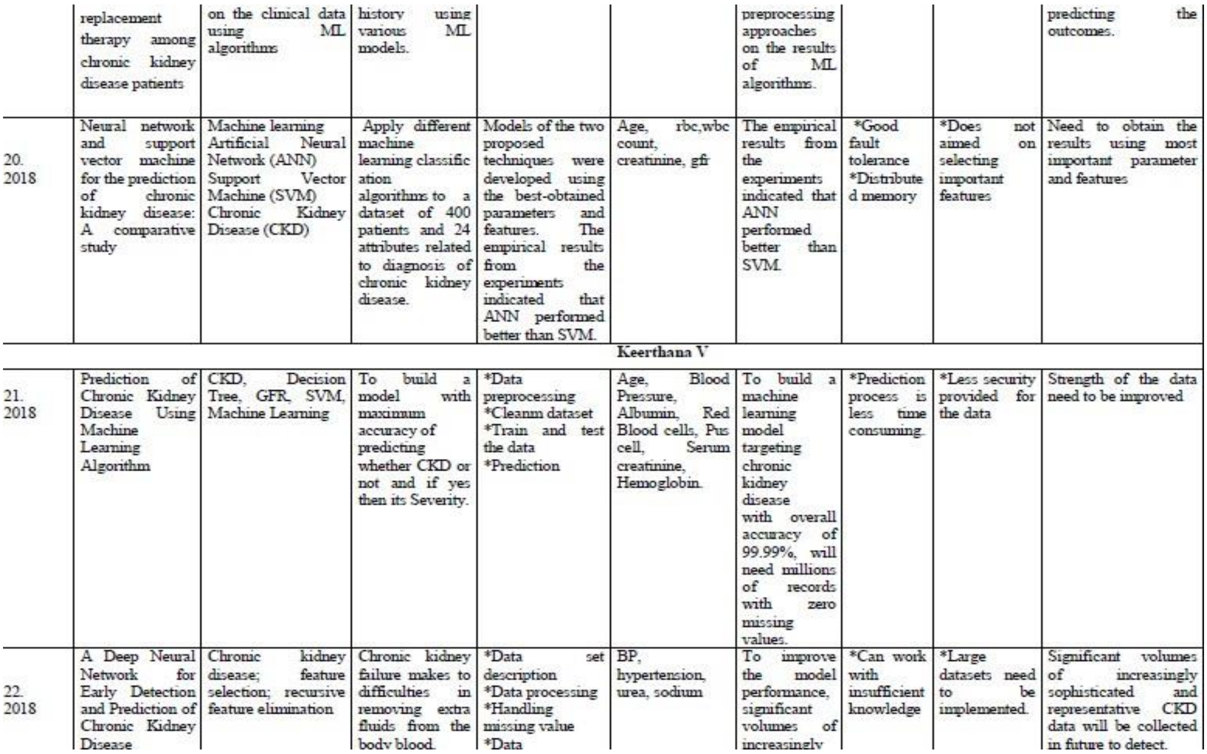


		Decision Tree (DT).							
3. 2022	A Deep Neural network for Early detection and Prediction of Chronic kidney disease	Regressive feature, elimination, support vector machine, machine learning	The proposed approach should be a useful tool for nephrologists in detecting CKD	*Data processing *Categorical data encoding *Data transformation *Outlier detection	Serum, Race, Historical data, genetic problems	The study concluded that it is more efficient in detecting CKD	*Execute feature engineerin g by itself	*It can handle only small datasets	To improve the model performance, significant volumes of increasingly sophisticated and representative CKD data will be collected in the future to detect disease severity
4. 2022	Early identification of CKD- a scoping review of the global publications	Chronic Kidney Disease (CKD), Machine Learning (ML), Support Vector Machine	Decisions on whether to screen for chronic kidney disease (CKD) or not remain contentious in nephrology. This study provides a global overview of early CKD identification efforts.	Data extracted from included studies focused on the following 4 themes: study population measurement methods, interventions used, and available policies.	Gender, bp, cholesterol, pulse rate	We identified 290 CKD screening and detection programs from 83 countries.	*Time saving process	This study has some limitations, including inability to use World Bank income grouping at time of the study	To explore the effect of ethnicity because most studies were not performed in homogeneous populations or did not report the racial composition of participants
5. 2022	Chronic Kidney Disease Predictionusing K-Means algorithm	Chronic kidney disease, k-means, Logistic regression Support vector machine	Elimination of disparities among kidney disease patients.	The best measure of kidney function is the glomerular filtration rate (GFR). To measure GFR is complica- ted in clinical practice, requiring substantial time and resources	*Albuminaria *GFR *Sodium	Stage 1-Normal Stage-2 Slightly damaged kidney Stage-3 Fully damaged kidney	1.Scales to large data sets. 2.Guarante es convergenc e.	1.Scaling with number of dimensions. 2.Clustering data of varying sizes and density.	To beat the performance of other classifiers using normalized dataset
					Tamilarasi S				
6. 2021	Diagnose of Chronic Kidney Disease by using	Naïve Bayes, Random Forest, eGFR, CKD,	The relationship of CKD to chronic kidney	Scan The Dataset, Calculate the	Blood pressure, sugar, bacteria,	Helps medical predicting the	*No need for anticoagula	*Protein loss through dailysate	It has to be enhanced based on the ground truth recommended



					model			overlapping.	
10. 2020	Detailed review of chronic kidney disease	Chronic kidney disease, Glomerular filtration rate, Albumin creatinine rate, Acute kidney injury, Cystatin C	With the help of various engineering techniques one can easily design controllers to assess as well as to prevent CKD permanently.	*Collect the dataset *Pre-process it *Apply machine learning algorithm *Predict the stage of CKD	RBC, WBC count, age, bp, sugar content	With the help of engineering techniques we can design predictive control systems for assessing as well as preventing CKD	*Simple to implement *More efficient	*More resources required	More accurate data mining techniques has to be implemented for better prediction
					Bhavani G				
11. 2020	The case for early identification and intervention of Chronic Kidney disease	Chronic kidney disease, Glomerular filtration rate, Albumin, creatinine rate	Participants identified strategies for screening, risk stratification, and treatment for early CKD and the key health system and economic factors for implementing these processes	*Input dataset *CKD Screening *Treatment	GFR, blood sample, urine sample	A fundamental justification for the early detection of CKD is the availability of evidence-based interventions to slow the progression of CKD and reduce its complications	*Easy to carry out *More throughput	*Need to work on additional features	Need to support systematic approaches for CKD screening
12. 2020	Risk Level Prediction of Chronic Kidney Disease using Neuro-Fuzzy and Hierarchical Clustering algorithm	Heatmap, clustering multivariant statistical analysis	The treatment can prevent or delay complications of decreased kidney function	A fusion of neural networks with fuzzy logic is generally defined as a system trained using a particular learning algorithm which is derived from neural network	BP, Glucose, Urine protein, urea	The results of the prediction showing the risk of any patient having CKD given the ten features.	*Data mining in healthcare detects fraud and abuse. *Help physicians to identify effective treatments	*Ethical, Legal and Social Issues. *Data Ownership issues.	Implement more bigdata oriented tools and technique which makes the process faster and effective



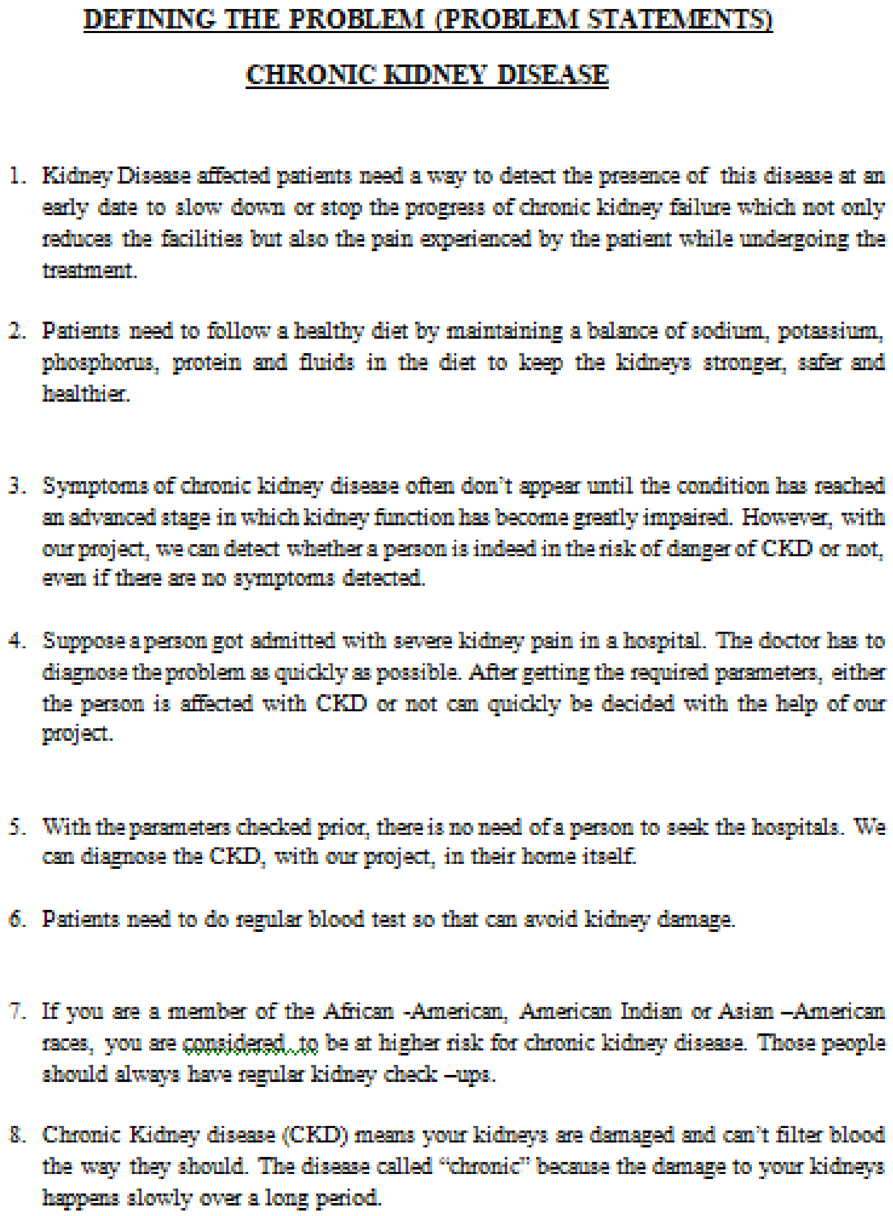


					Kanimozhi D				
16. 2019	Chronic Kidney Disease Prediction Using Machine Learning Techniques	Chronic kidney disease Decision Tree -Support Vector Machine Logistic Regression and Bagging Ensemble methods	To develop and validate a predictive model for the prediction of chronic kidney disease.	*CKD dataset *Apply logistic regression and decision tree algorithm *Bagging ensemble technique *Result	Glucose, BP, Albumin, Creatinine	The machine learning classifiers used in two stages, at first the logistic regression classifier is used to predict the results	*Easy to implement *Accurate	*Not efficient	The model can be further turned by applying feature selection methods to increase the performance of the prediction
17. 2019	Chronic Kidney Disease Prediction using Machine Learning Models	Chronic Kidney Disease, Decision Tree, Machine Learning, Random Forest, Support Vactors.	Presents evidences of early identification and care of CKD can improve the quality of the patients life.	*Create dataset *Pre-processing *Apply decision and support vector machine.	Glucose, BP, RBC, WBC, Albumin	Models has been constructed using training data set instances which is 70% of original CKD data set. Constructed models have been validated using test data.	*Efficient functioning *Simple to implement	*Not compared the data at the time of execution	The comparison must be done based on the time of execution and feature set selection
18. 2019	Detection of Chronic Kidney Disease Using Machine Learning Algorithms with Least Number of Predictors	Chronic kidney disease, Random forest, Gradient boosting, Logistic Regression, Support vector machines, Machine learning.	The target is to diagnose the CKD using different intelligent techniques.	*Data preprocessing *Missing values *Data reduction *Modelling *Result	Age, sex, serum, creatinine, albumin	The experiments are conducted using Python 3.3 programming language through the Jupyter Notebook	*Effective method	*Disassociation between classes may occur	It is aimed to validate the results by using bigdata sets or compare the result using another dataset that contains the same features.
19. 2019	Using machine learning models to predict the initiation of renal replacement therapy among chronic kidney disease patients	A novel approach of screening CKD patients to predict the chances of future RRT based on the clinical data using ML algorithms	The objective of our study is to develop a screening tool based on disease history using various ML models.	*Study design *Data source *Petient selection *Outcome definition	Age, serum, urea, creatinine, albumin, gfr	They, tested the effect of all the implemented data preprocessing approaches on the results of ML algorithms.	*Easy data aquisation	*High error prone *Time consuming	Future scope lies in coming up with a prediction model that would factor in the more clinical data in predicting the outcomes.
20. 2018	Neural network and support vector machine for the prediction of chronic kidney disease: A comparative study	Machine learning Artificial Neural Network (ANN) Support Vector Machine (SVM) Chronic Kidney Disease (CKD)	Apply different machine learning classification algorithms to a dataset of 400 patients and 24 attributes related to diagnosis of chronic kidney disease.	Models of the two proposed techniques were developed using the best-obtained parameters and features. The empirical results from the experiments indicated that ANN performed better than SVM.	Age, rbc,wbc count, creatinine, gfr	The empirical results from the experiments indicated that ANN performed better than SVM.	*Good fault tolerance *Distributed memory	*Does not aimed on selecting important features	Need to obtain the results using most important parameter and features
					Keerthana V				
21. 2018	Prediction of Chronic Kidney Disease Using Machine Learning Algorithm	CKD, Decision Tree, GFR, SVM, Machine Learning	To build a model with maximum accuracy of predicting whether CKD or not and if yes then its Severity.	*Data preprocessing *Cleanm dataset *Train and test the data *Prediction	Age, Blood Pressure, Albumin, Red Blood cells, Pus cell, Serum creatinine, Hemoglobin.	To build a machine learning model targeting chronic kidney disease with overall accuracy of 99.99%, will need millions of records with zero missing values.	*Prediction process is less time consuming.	*Less security provided for the data	Strength of the data need to be improved
22. 2018	A Deep Neural Network for Early Detection and Prediction of Chronic Kidney Disease	Chronic kidney disease; feature selection; recursive feature elimination	Chronic kidney failure makes to difficulties in removing extra fluids from the body blood.	*Data set description *Data processing *Handling missing value *Data	BP, hypertension, urea, sodium	To improve the model performance, significant volumes of increasingly	*Can work with insufficient knowledge	*Large datasets need to be implemented.	Significant volumes of increasingly sophisticated and representative CKD data will be collected in future to detect.

Tables 2.2.1 Literature Survey

2.3 Problem Statement Definition:

- The first step in the problem-solving process is to determine what the problem actually is. This is an important step because you can waste time solving the wrong problem. Do not assume automatically you know what the problem is, because it may not be apparent.
- The problem statement is a structured set of statements that describe the purpose of an effort in terms of what problem it's trying to solve.



DEFINING THE PROBLEM (PROBLEM STATEMENTS)

CHRONIC KIDNEY DISEASE

1. Kidney Disease affected patients need a way to detect the presence of this disease at an early date to slow down or stop the progress of chronic kidney failure which not only reduces the facilities but also the pain experienced by the patient while undergoing the treatment.
2. Patients need to follow a healthy diet by maintaining a balance of sodium, potassium, phosphorus, protein and fluids in the diet to keep the kidneys stronger, safer and healthier.
3. Symptoms of chronic kidney disease often don't appear until the condition has reached an advanced stage in which kidney function has become greatly impaired. However, with our project, we can detect whether a person is indeed in the risk of danger of CKD or not, even if there are no symptoms detected.
4. Suppose a person got admitted with severe kidney pain in a hospital. The doctor has to diagnose the problem as quickly as possible. After getting the required parameters, either the person is affected with CKD or not can quickly be decided with the help of our project.
5. With the parameters checked prior, there is no need of a person to seek the hospitals. We can diagnose the CKD, with our project, in their home itself.
6. Patients need to do regular blood test so that can avoid kidney damage.
7. If you are a member of the African -American, American Indian or Asian –American races, you are considered to be at higher risk for chronic kidney disease. Those people should always have regular kidney check –ups.
8. Chronic Kidney disease (CKD) means your kidneys are damaged and can't filter blood the way they should. The disease called "chronic" because the damage to your kidneys happens slowly over a long period.

3. IDEATION & PROPOSED SOLUTION

3.1 Empathy Map Canvas:

An empathy map is a collaborative visualization used to articulate what we know about a particular type of user. It externalizes knowledge about users in order to 1) create a shared understanding of user needs, and 2) aid in decision making.

An empathy map helps to map what a design team knows about the potential audience. This tool helps to understand the reason behind some actions a user takes deeply. This tool helps build Empathy towards users and helps design teams shift focus from the product to the users who are going to use the product.

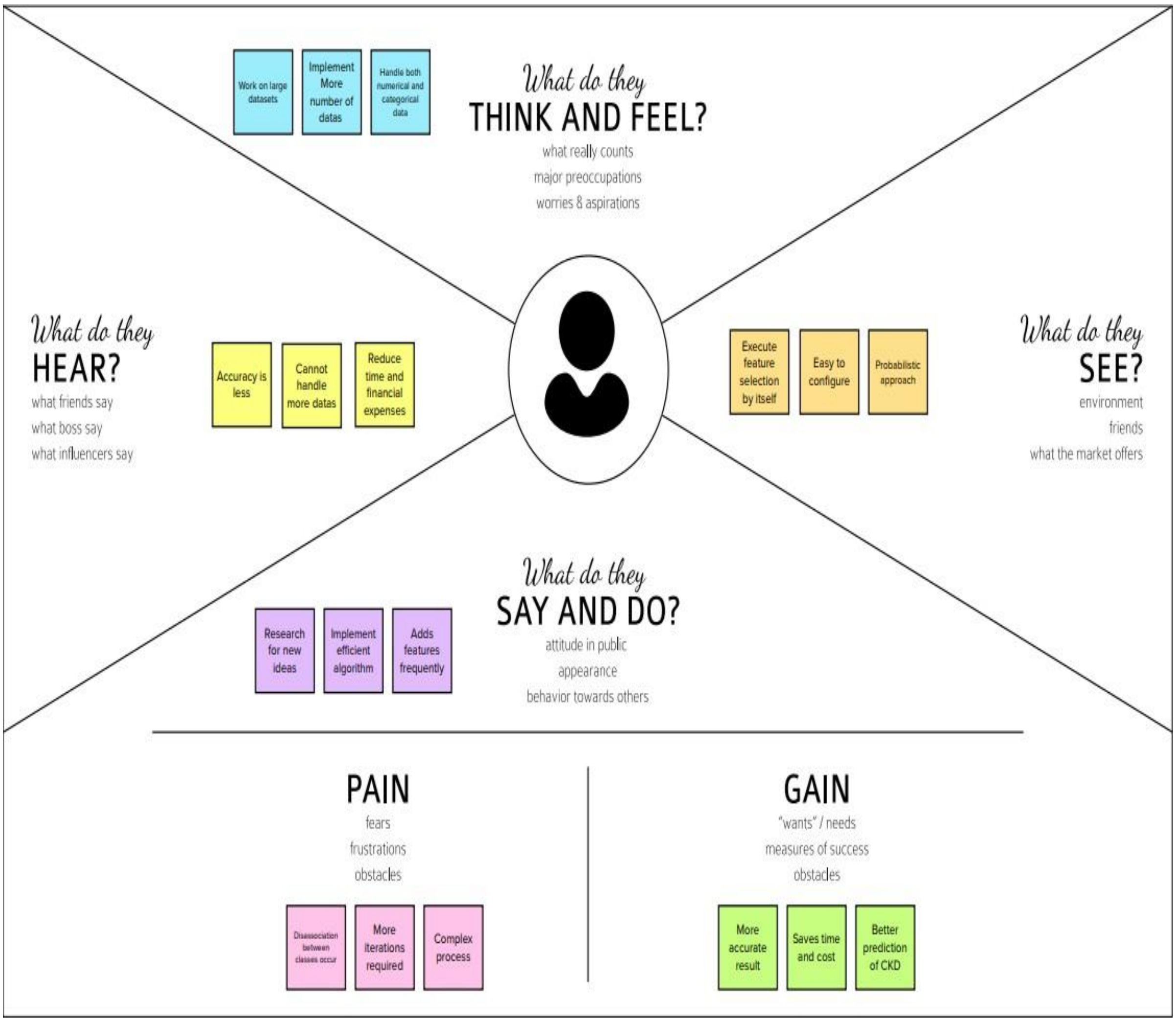


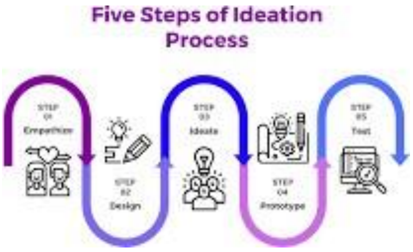
Fig 3.1.1 Empathy Map Canvas

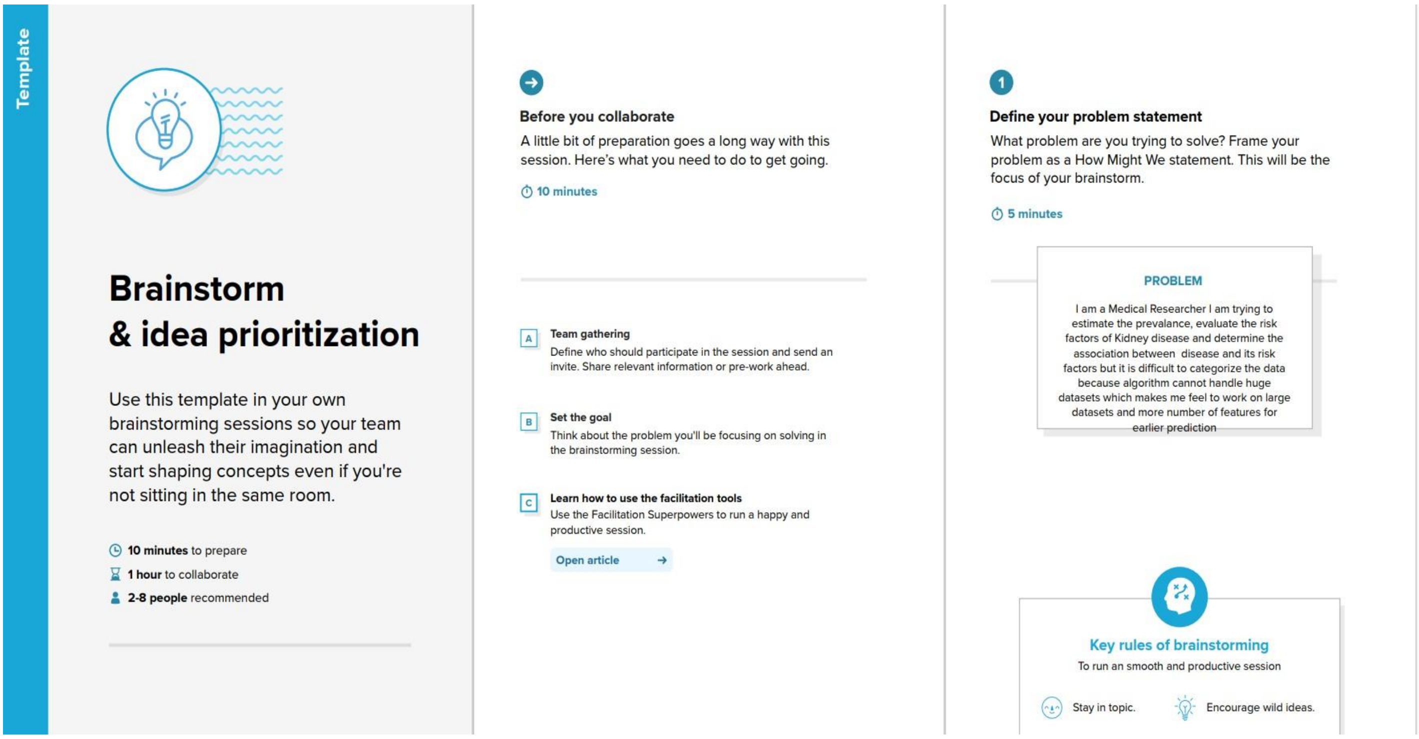
3.2 Ideation and Brainstorming:

- Brainstorming is an activity that will help you generate more innovative ideas. It's one of many methods of ideation—the process of coming up with new ideas—and it's core to the design thinking process.

- Brainstorming refers to a problem-solving technique used by teams or individuals. In this process, participants generate various ideas or solutions, then begin discussing and narrowing them down to the best options.

- Ideation is often closely related to the practice of brainstorming, a specific technique that is utilized to generate new ideas. A principal difference between ideation and brainstorming is that ideation is commonly more thought of as being an individual pursuit, while brainstorming is almost always a group activity.





Template

Brainstorm & idea prioritization

Use this template in your own brainstorming sessions so your team can unleash their imagination and start shaping concepts even if you're not sitting in the same room.

10 minutes to prepare
1 hour to collaborate
2-8 people recommended

Before you collaborate

A little bit of preparation goes a long way with this session. Here's what you need to do to get going.

10 minutes

A **Team gathering**
Define who should participate in the session and send an invite. Share relevant information or pre-work ahead.

B **Set the goal**
Think about the problem you'll be focusing on solving in the brainstorming session.

C **Learn how to use the facilitation tools**
Use the Facilitation Superpowers to run a happy and productive session.

Open article →

1 **Define your problem statement**

What problem are you trying to solve? Frame your problem as a How Might We statement. This will be the focus of your brainstorm.

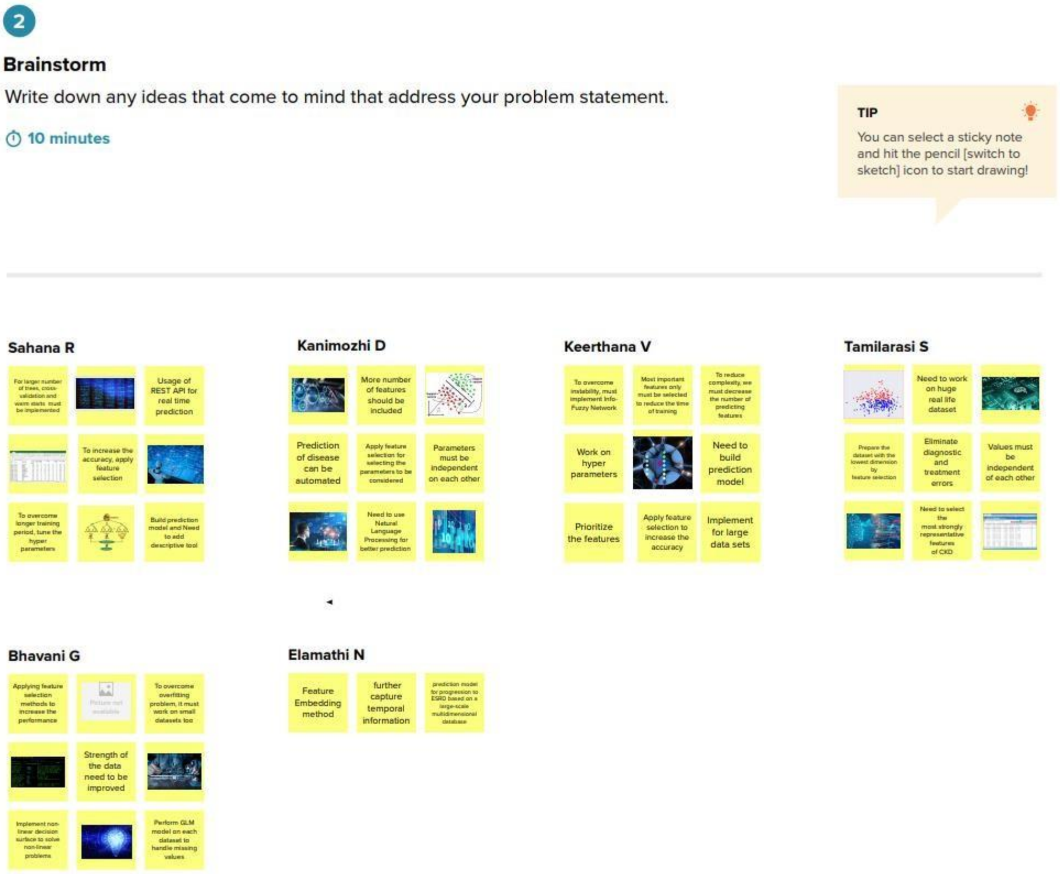
5 minutes

PROBLEM

I am a Medical Researcher I am trying to estimate the prevalance, evaluate the risk factors of Kidney disease and determine the association between disease and its risk factors but it is difficult to categorize the data because algorithm cannot handle huge datasets which makes me feel to work on large datasets and more number of features for earlier prediction

Key rules of brainstorming
To run an smooth and productive session

Stay in topic.   Encourage wild ideas.



2 **Brainstorm**

Write down any ideas that come to mind that address your problem statement.

10 minutes

TIP
You can select a sticky note and hit the pencil [switch to sketch] icon to start drawing!

Sahana R

For larger number of trees, cross-validation and warm starts must be implemented

Usage of REST API for real time prediction

To increase the accuracy, apply feature selection

To overcome longer training period, tune the hyper parameters

Build prediction model and Need to add descriptive tool

Kanimozhi D

More number of features should be included

Prediction of disease can be automated

Apply feature selection for selecting the parameters to be considered

Parameters must be independent on each other

Need to use Natural Language Processing for better prediction

Keerthana V

To overcome instability, must implement Info-Fuzzy Network

Most important features only must be selected to reduce the time of training

To reduce complexity, we must decrease the number of predicting features

Work on hyper parameters

Need to build prediction model

Prioritize the features

Apply feature selection to increase the accuracy

Implement for large data sets

Tamilarasi S

Need to work on huge real life dataset

Prepare the dataset with the lowest dimension by feature selection

Eliminate diagnostic and treatment errors

Values must be independent of each other

Need to select the most strongly representative features of CKD

Bhavani G

Applying feature selection methods to increase the performance

Picture not available

To overcome overfitting problem, it must work on small datasets too

Strength of the data need to be improved

Implement non-linear decision surface to solve non-linear problems

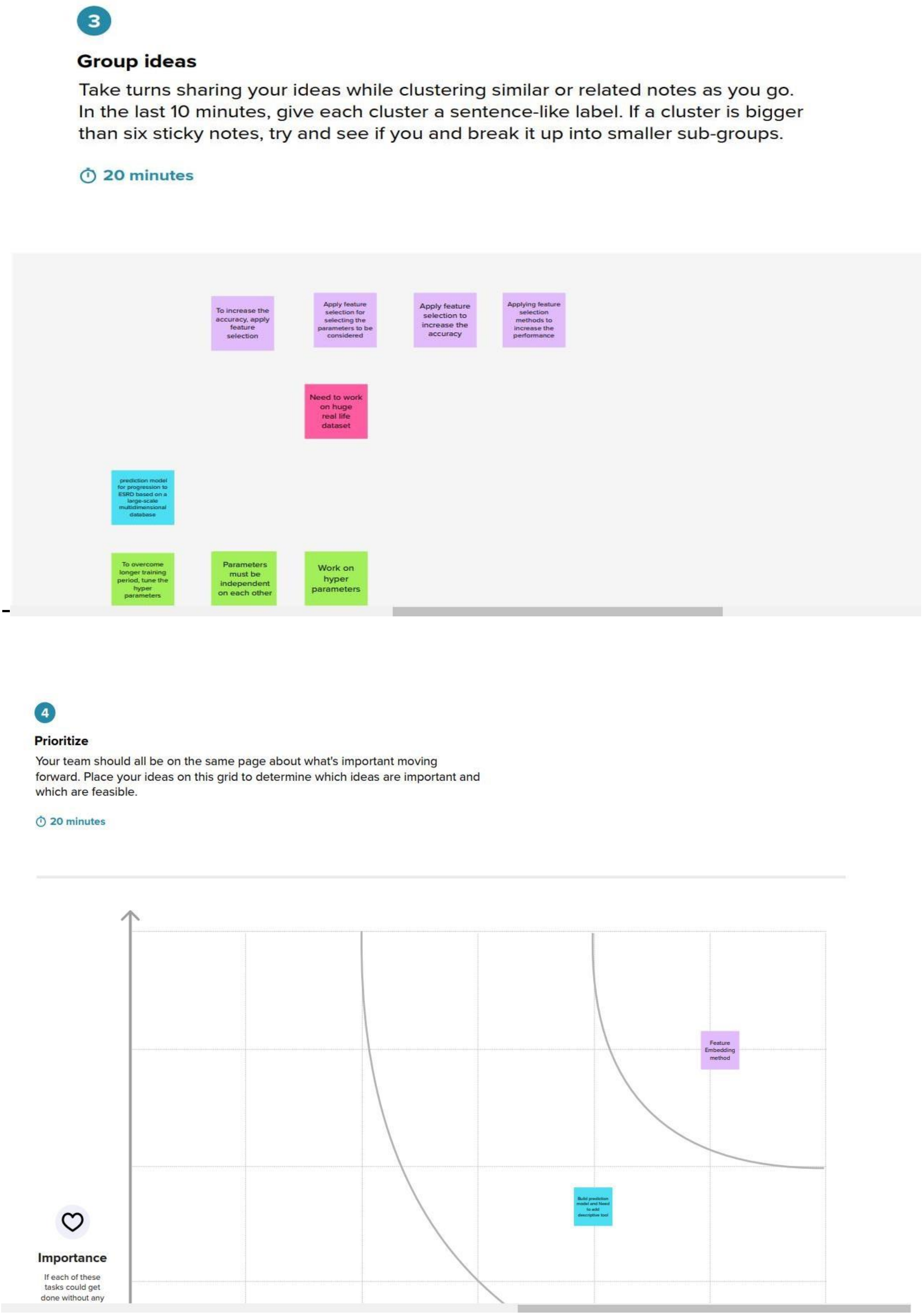
Perform GLM model on each dataset to handle missing values

Elamathi N

Feature Embedding method

further capture temporal information

prediction model for progression to ESRD based on a large-scale multidimensional database



3

Group ideas

Take turns sharing your ideas while clustering similar or related notes as you go. In the last 10 minutes, give each cluster a sentence-like label. If a cluster is bigger than six sticky notes, try and see if you and break it up into smaller sub-groups.

20 minutes

To increase the accuracy, apply feature selection

Apply feature selection for selecting the parameters to be considered

Apply feature selection to increase the accuracy

Applying feature selection methods to increase the performance

Need to work on huge real life dataset

prediction model for progression to ESRD based on a large-scale multidimensional database

To overcome longer training period, tune the hyper parameters

Parameters must be independent on each other

Work on hyper parameters

4

Prioritize

Your team should all be on the same page about what's important moving forward. Place your ideas on this grid to determine which ideas are important and which are feasible.

20 minutes

Feature Embedding method

Build prediction model and Need to add descriptive tool

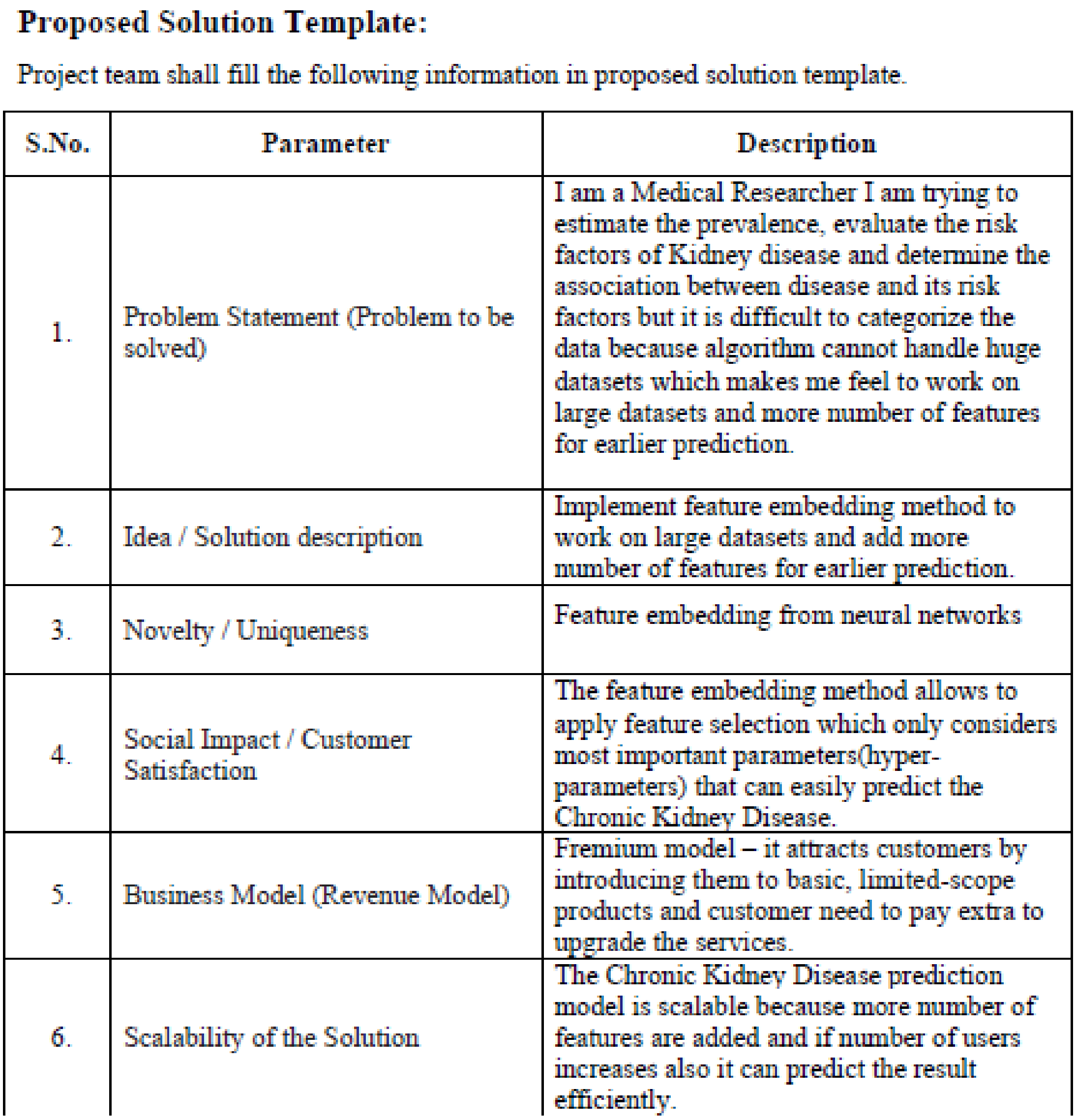
If each of these tasks could get done without any

Fig 3.2.1 Brainstorming

3.3 Proposed Solution:

- The purpose of this tool is to provide a structured process for identifying a problem, understanding the root causes, ascertaining solution steps, and progress monitoring.

- With a solution template, you can organize development content that you want to reuse for customer-specific solutions. Solution templates enable you to easily start the development of customer-specific solutions, for example, for a specific industry.



Proposed Solution Template:

Project team shall fill the following information in proposed solution template.

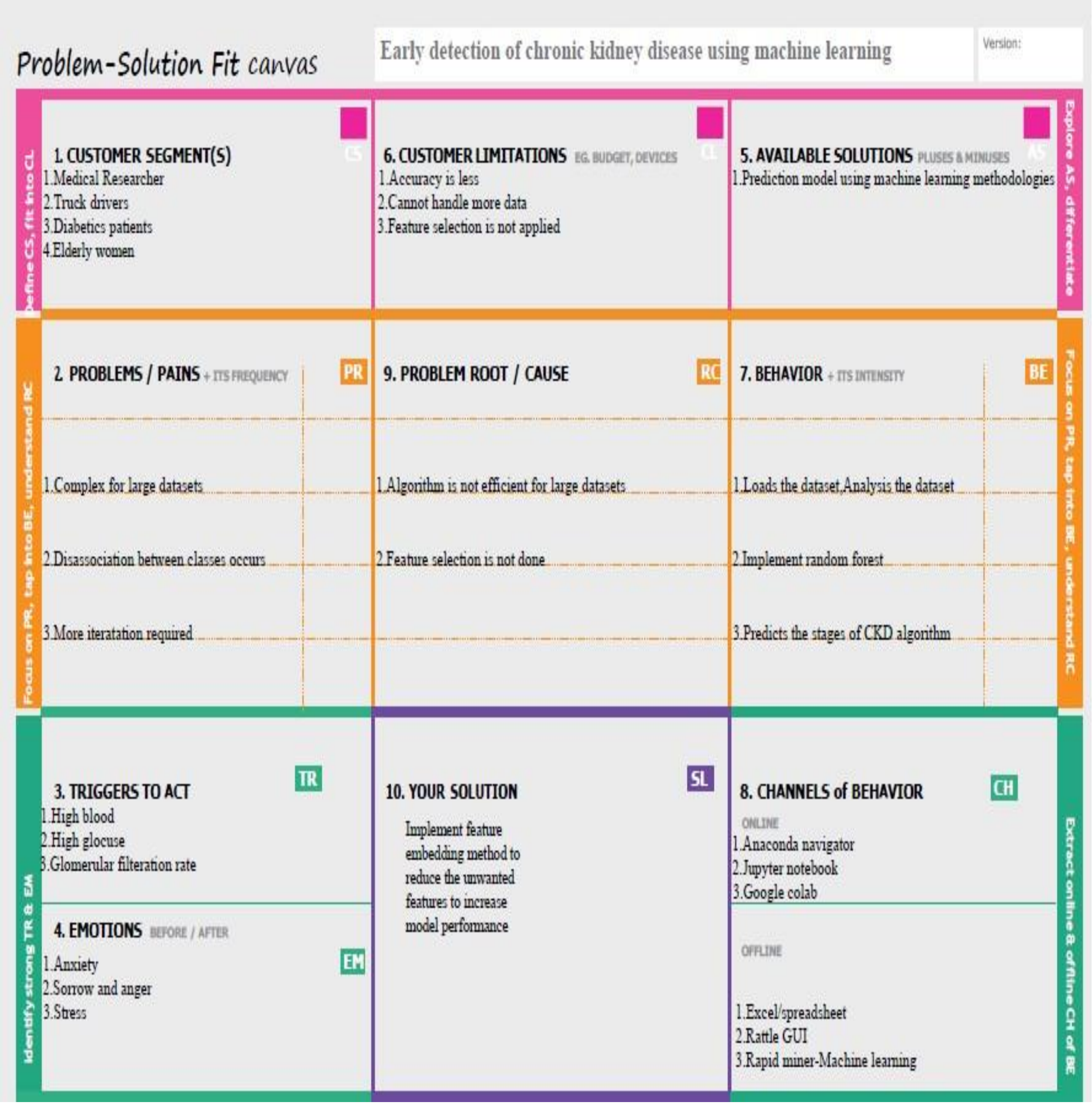
S.No.	Parameter	Description
1.	Problem Statement (Problem to be solved)	I am a Medical Researcher I am trying to estimate the prevalence, evaluate the risk factors of Kidney disease and determine the association between disease and its risk factors but it is difficult to categorize the data because algorithm cannot handle huge datasets which makes me feel to work on large datasets and more number of features for earlier prediction.
2.	Idea / Solution description	Implement feature embedding method to work on large datasets and add more number of features for earlier prediction.
3.	Novelty / Uniqueness	Feature embedding from neural networks
4.	Social Impact / Customer Satisfaction	The feature embedding method allows to apply feature selection which only considers most important parameters(hyper-parameters) that can easily predict the Chronic Kidney Disease.
5.	Business Model (Revenue Model)	Fremium model – it attracts customers by introducing them to basic, limited-scope products and customer need to pay extra to upgrade the services.
6.	Scalability of the Solution	The Chronic Kidney Disease prediction model is scalable because more number of features are added and if number of users increases also it can predict the result efficiently.

Table 3.3.1 Proposed Solution

3.4 Problem Solution Fit:

The Problem-Solution Fit simply means that you have found a problem with your customer and that the solution you have realized for it actually solves the customer's problem.

Problem-Solution Fit - this occurs when you have evidence that customers care about certain jobs, pains, and gains. At this stage you've proved the existence of a problem and have designed a value proposition that addresses your customers' jobs, pains and gains.



Problem-Solution Fit canvas

Early detection of chronic kidney disease using machine learning

Version:

Define CS, fit into CL
Explore AS, differentiate
Focus on PR, tap into BE, understand RC
Focus on PR, tap into BE, understand RC
Identify strong TR & EM
Extract online & offline CH of BE
CS
CL
AS
PR
RC
BE
TR
SL
CH
EM

1. CUSTOMER SEGMENT(S)
1.Medical Researcher
2.Truck drivers
3.Diabetics patients
4.Elderly women

6. CUSTOMER LIMITATIONS EG. BUDGET, DEVICES
1.Accuracy is less
2.Cannot handle more data
3.Feature selection is not applied

5. AVAILABLE SOLUTIONS PLUSES & MINUSES
1.Prediction model using machine learning methodologies

2. PROBLEMS / PAINS + ITS FREQUENCY
1.Complex for large datasets
2.Disassociation between classes occurs
3.More iteratation required

9. PROBLEM ROOT / CAUSE
1.Algorithm is not efficient for large datasets
2.Feature selection is not done

7. BEHAVIOR + ITS INTENSITY
1.Loads the dataset,Analysis the dataset
2.Implement random forest
3.Predicts the stages of CKD algorithm

3. TRIGGERS TO ACT
1.High blood
2.High glocuse
3.Glomerular filteration rate

4. EMOTIONS BEFORE / AFTER
1.Anxiety
2.Sorrow and anger
3.Stress

10. YOUR SOLUTION
Implement feature embedding method to reduce the unwanted features to increase model performance

8. CHANNELS of BEHAVIOR
ONLINE
1.Anaconda navigator
2.Jupyter notebook
3.Google colab
OFFLINE
1.Excel/spreadsheet
2.Rattle GUI
3.Rapid miner-Machine learning

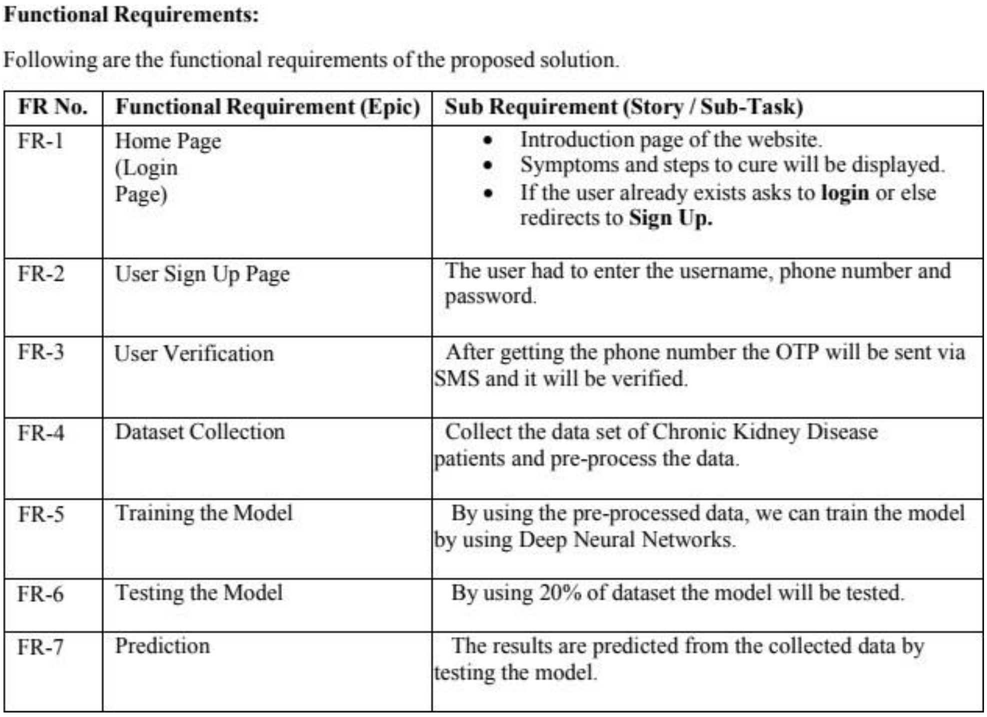
Fig 3.4.1 Problem Solution Fit

4. REQUIREMENT ANALYSIS

4.1 Functional requirement:

- Solution Requirements are identified before the technical solution is selected and/or designed. They describe the characteristics of a solution (functional and non-functional) that meet business requirements and stakeholder requirements.

- A solution requirement is aimed at the concerns of the people who will build and deliver the solution. It tells those people what the functional and non-functional requirements for the solution will be and how the solution will deliver on the business and stakeholder requirements. Solution Requirements – Describe the features, functions, and characteristics of a product, service, or result that will meet the business and stakeholder requirements.



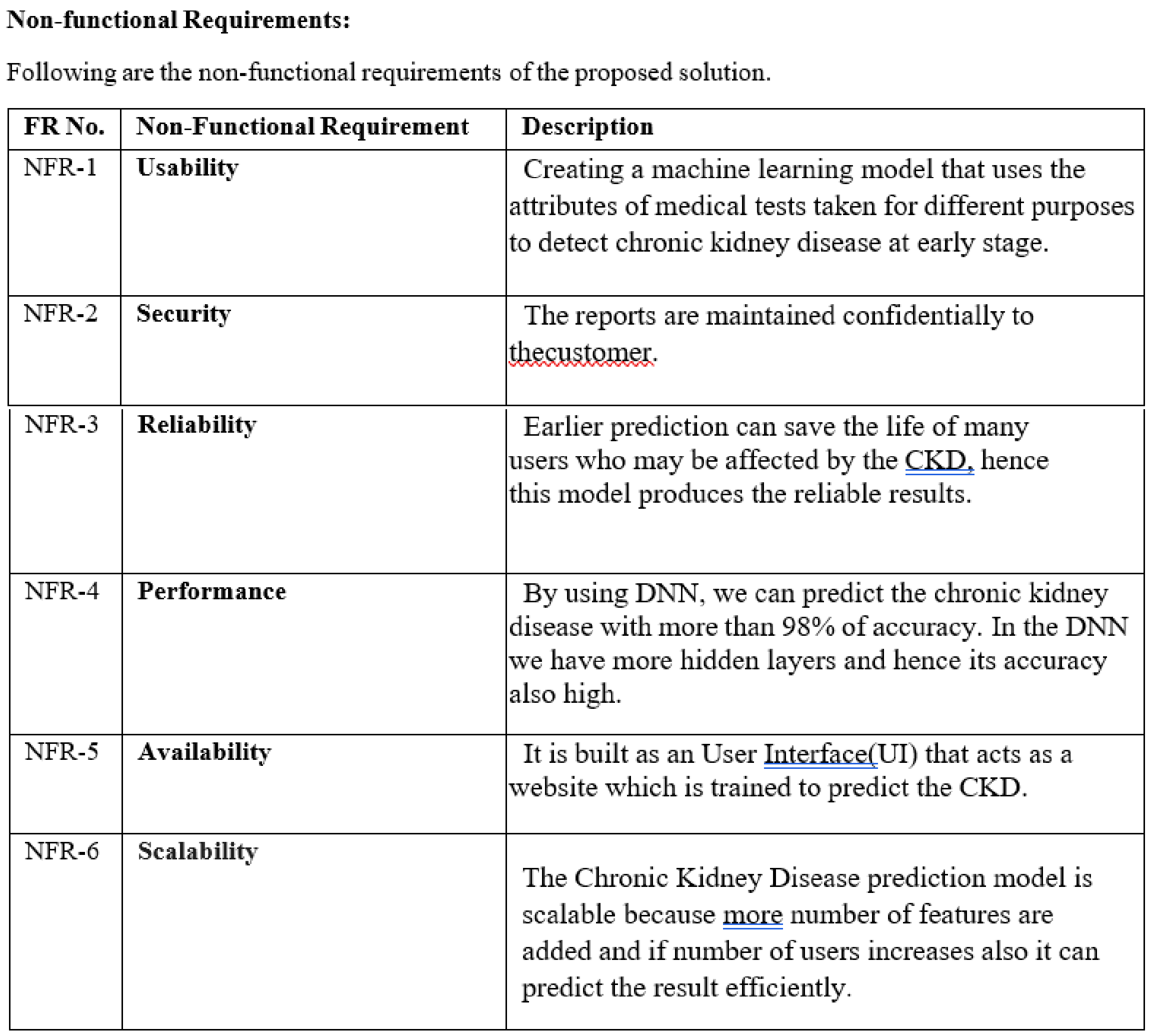
Functional Requirements:

Following are the functional requirements of the proposed solution.

FR No.	Functional Requirement (Epic)	Sub Requirement (Story / Sub-Task)
FR-1	Home Page (Login Page)	• Introduction page of the website. • Symptoms and steps to cure will be displayed. • If the user already exists asks to **login** or else redirects to **Sign Up.**
FR-2	User Sign Up Page	The user had to enter the username, phone number and password.
FR-3	User Verification	After getting the phone number the OTP will be sent via SMS and it will be verified.
FR-4	Dataset Collection	Collect the data set of Chronic Kidney Disease patients and pre-process the data.
FR-5	Training the Model	By using the pre-processed data, we can train the model by using Deep Neural Networks.
FR-6	Testing the Model	By using 20% of dataset the model will be tested.
FR-7	Prediction	The results are predicted from the collected data by testing the model.

Table 4.1.1 Functional Requirement

4.2 Non-Functional Requirement:



Non-functional Requirements:

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

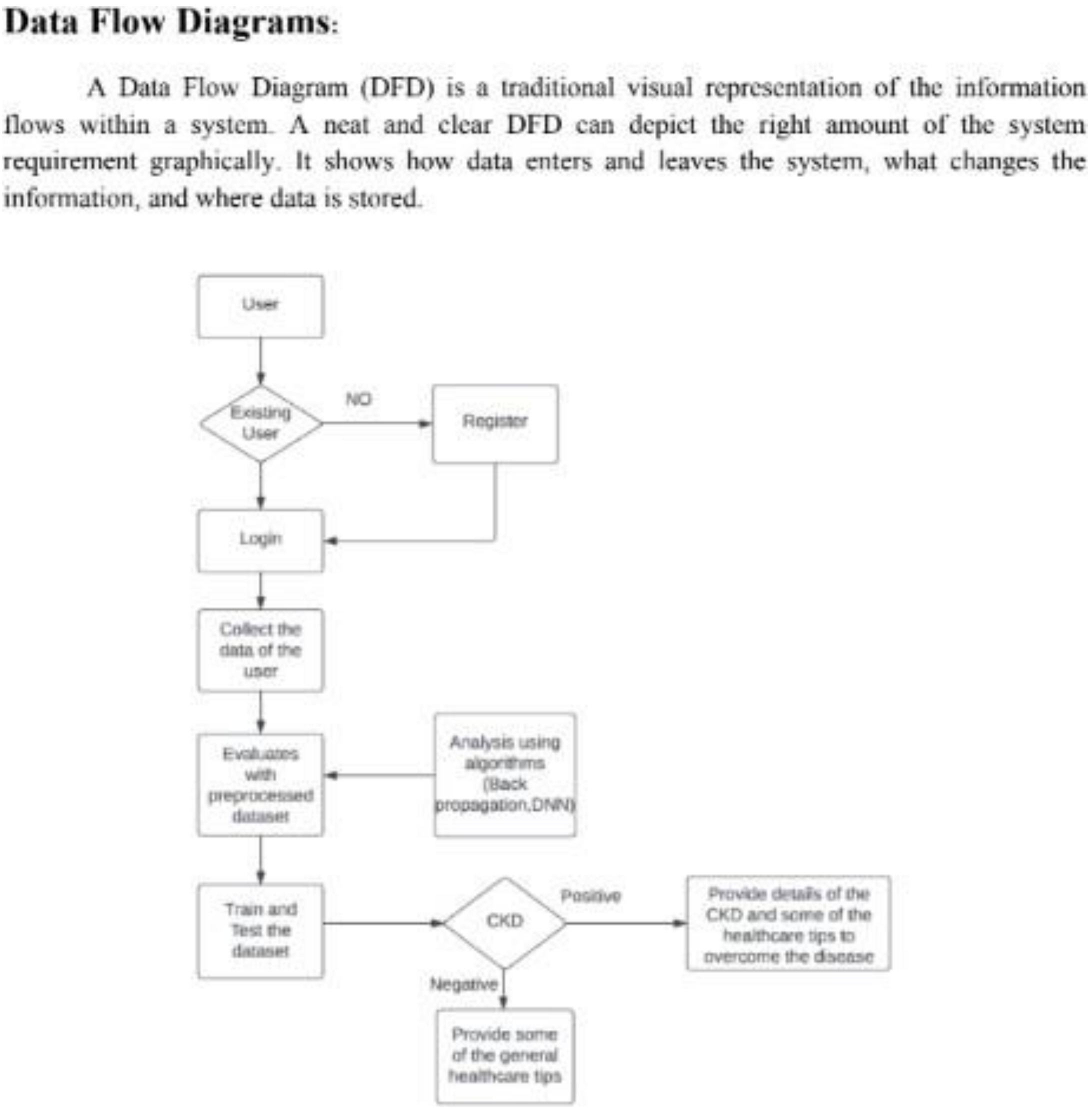
FR No.	Non-Functional Requirement	Description
NFR-1	**Usability**	Creating a machine learning model that uses the attributes of medical tests taken for different purposes to detect chronic kidney disease at early stage.
NFR-2	**Security**	The reports are maintained confidentially to thecustomer.
NFR-3	**Reliability**	Earlier prediction can save the life of many users who may be affected by the CKD, hence this model produces the reliable results.
NFR-4	**Performance**	By using DNN, we can predict the chronic kidney disease with more than 98% of accuracy. In the DNN we have more hidden layers and hence its accuracy also high.
NFR-5	**Availability**	It is built as an User Interface(UI) that acts as a website which is trained to predict the CKD.
NFR-6	**Scalability**	The Chronic Kidney Disease prediction model is scalable because more number of features are added and if number of users increases also it can predict the result efficiently.

Table 4.2.1 Non-Functional Requirement

5. PROJECT DESIGN

5.1 Data Flow Diagram:

- A Data Flow Diagram (DFD) is a graphical representation of the "flow" of data through an information system(as shown on the DFD flow chart Figure 5), modeling its process aspects. Often it is a preliminary step used to create an overview of the system that can later be elaborated.



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.

User
Existing User
NO
Register
Login
Collect the data of the user
Evaluates with preprocessed dataset
Analysis using algorithms (Back propagation,DNN)
Train and Test the dataset
CKD
Positive
Provide details of the CKD and some of the healthcare tips to overcome the disease
Negative
Provide some of the general healthcare tips

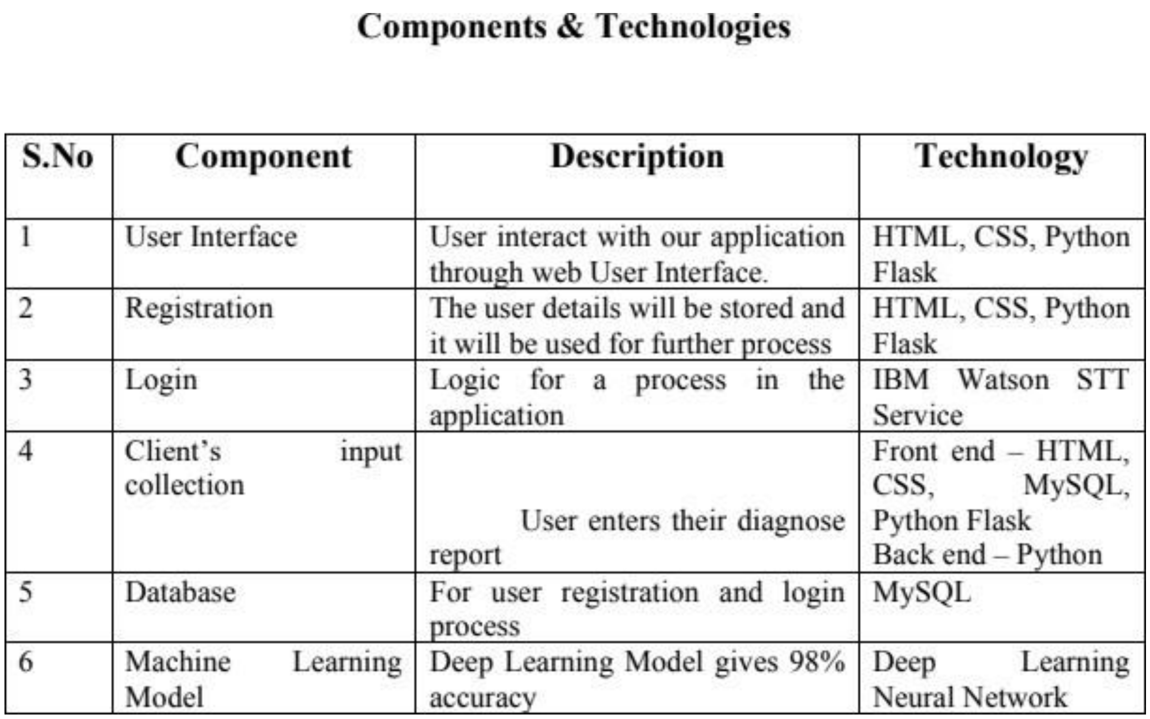
Fig 5.1.1 Data Flow Diagram

5.2 Solution and Technical Architecture:

- A solutions architect creates the overall technical vision for a specific solution to a business problem. A solutions architect creates the overall technical vision for a specific solution to a business problem. They design, describe, and manage the solution.

- Technology Architecture describes the logical software and hardware capabilities that are required to support the deployment of business, data, and application services. This includes IT infrastructure, middleware, networks, communications, processing, standards, etc.

- Technology architecture deals with the deployment of application components on technology components. A standard set of predefined technology components is provided in order to represent servers, network, workstations.



Components & Technologies

S.No	Component	Description	Technology
1	User Interface	User interact with our application through web User Interface.	HTML, CSS, Python Flask
2	Registration	The user details will be stored and it will be used for further process	HTML, CSS, Python Flask
3	Login	Logic for a process in the application	IBM Watson STT Service
4	Client's input collection	User enters their diagnose report	Front end – HTML, CSS, MySQL, Python Flask Back end – Python
5	Database	For user registration and login process	MySQL
6	Machine Learning Model	Deep Learning Model gives 98% accuracy	Deep Learning Neural Network

Table 5.2.1 Components & Technologies

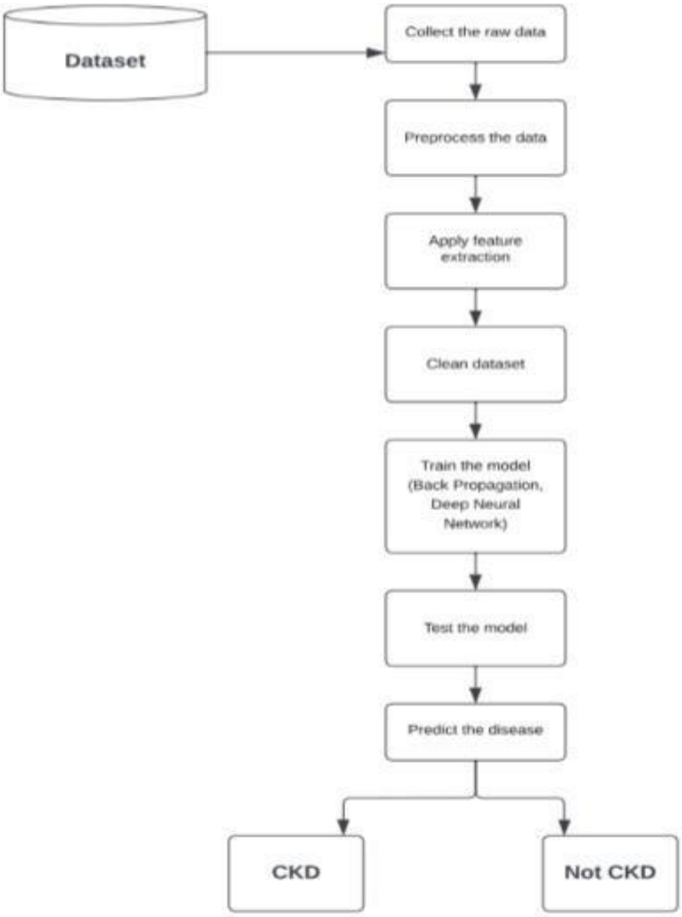
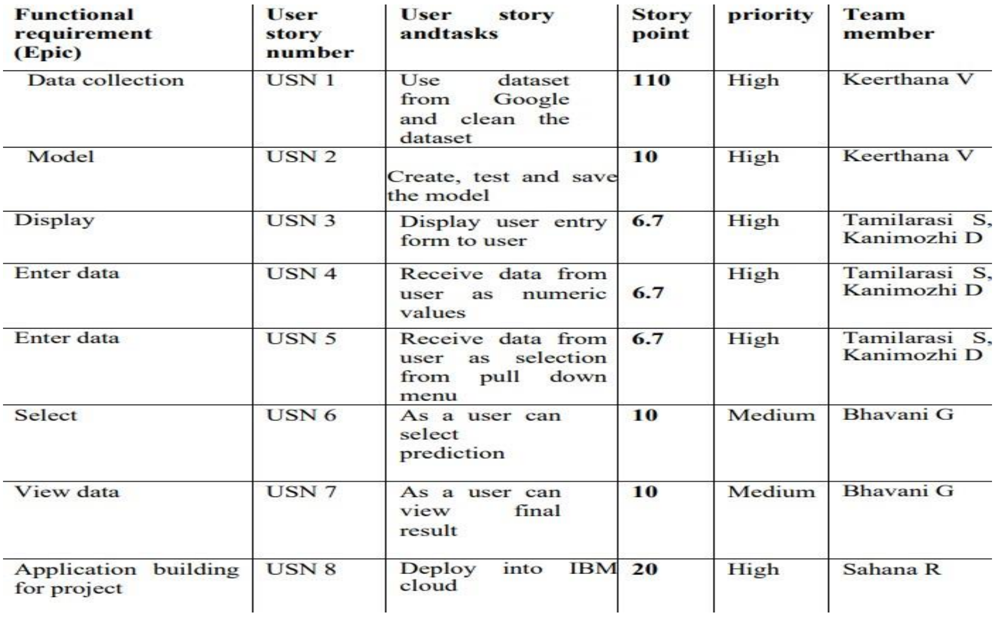


Fig 5.2.1 Technical Architecture

5.3 User Stories:



Functional requirement (Epic)	User story number	User story andtasks	Story point	priority	Team member
Data collection	USN 1	Use dataset from Google and clean the dataset	110	High	Keerthana V
Model	USN 2	Create, test and save the model	10	High	Keerthana V
Display	USN 3	Display user entry form to user	6.7	High	Tamilarasi S, Kanimozhi D
Enter data	USN 4	Receive data from user as numeric values	6.7	High	Tamilarasi S, Kanimozhi D
Enter data	USN 5	Receive data from user as selection from pull down menu	6.7	High	Tamilarasi S, Kanimozhi D
Select	USN 6	As a user can select prediction	10	Medium	Bhavani G
View data	USN 7	As a user can view final result	10	Medium	Bhavani G
Application building for project	USN 8	Deploy into IBM cloud	20	High	Sahana R

Table 5.3.1 User Stories

Customer Journey Map:

- During the Design Phase II we have done Customer journey map, Data flow diagram & user stories, Solution Requirement and Technology Architecture. Let us see in detail each activity.

- A customer journey map is a visual storyline of every engagement a customer has with a service, brand, or product. The creation of a journey map puts the organization directly in the mind of the consumer, so they can see and understand their customer's processes, needs, and perceptions.

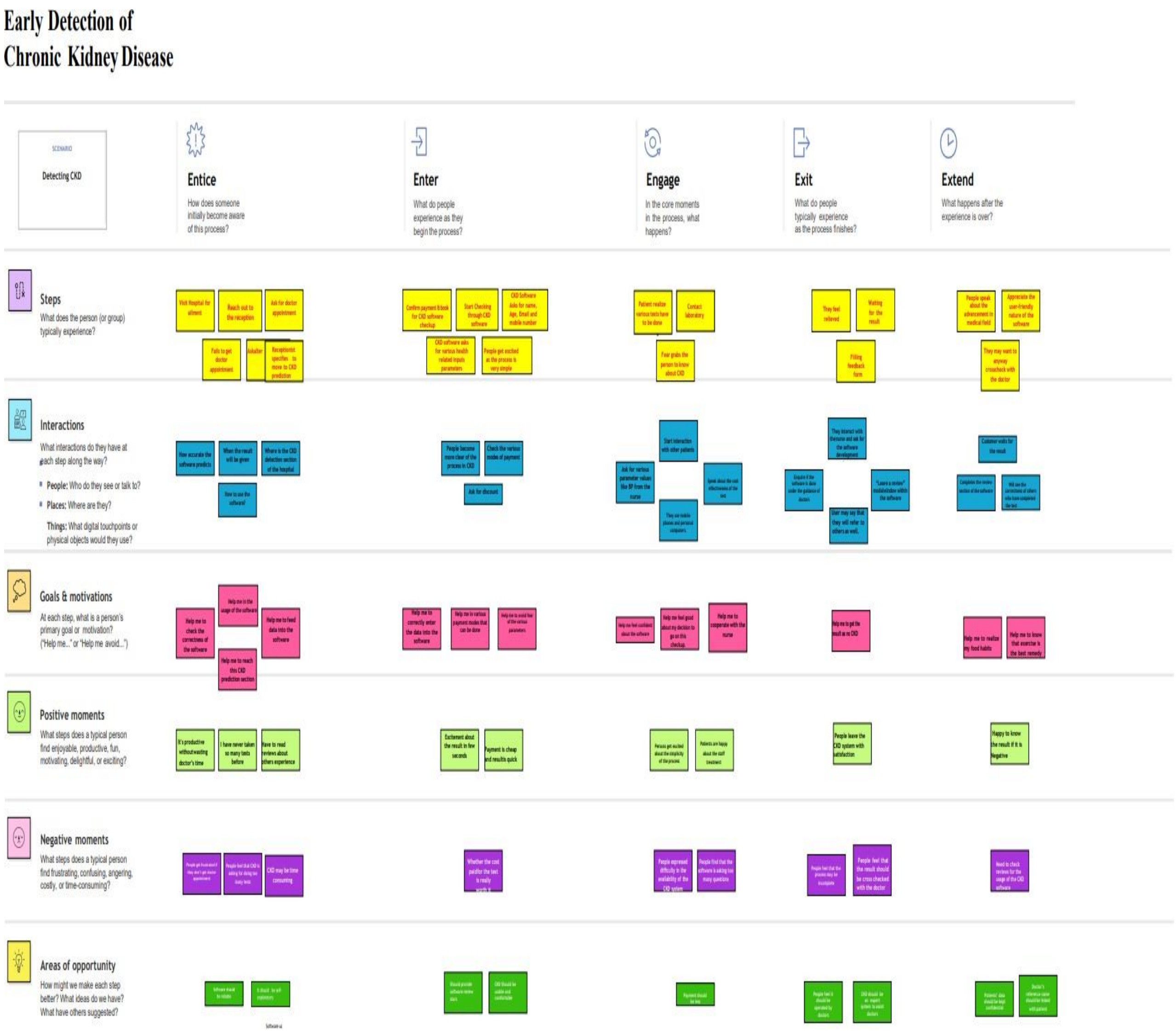

Early Detection of Chronic Kidney Disease
SCENARIO
Detecting CKD
Entice
How does someone initially become aware of this process?
Enter
What do people experience as they begin the process?
Engage
In the core moments in the process, what happens?
Exit
What do people typically experience as the process finishes?
Extend
What happens after the experience is over?
Steps
What does the person (or group) typically experience?
Visit Hospital for ailment
Reach out to the reception
Ask for doctor appointment
Fails to get doctor appointment
Askalter
Receptionist specifies to move to CKD prediction
Confirm payment & book for CKD software checkup
Start Checking through CKD software
CKD Software Asks for name, Age, Email and mobile number
CKD software asks for various health related inputs parameters
People get excited as the process is very simple
Patient realize various tests have to be done
Contact laboratory
Fear grabs the person to know about CKD
They feel relieved
Waiting for the result
Filling feedback form
People speak about the advancement in medical field
Appreciate the user-friendly nature of the software
They may want to anyway crosscheck with the doctor
Interactions
What interactions do they have at each step along the way?
People: Who do they see or talk to?
Places: Where are they?
Things: What digital touchpoints or physical objects would they use?
How accurate the software predicts
When the result will be given
Where is the CKD detection section of the hospital
How to use the software?
People become more clear of the process in CKD
Check the various modes of payment
Ask for discount
Start interaction with other patients
Ask for various parameter values like BP from the nurse
Speak about the cost effectiveness of the test
They use mobile phones and personal computers.
They interact with the nurse and ask for the software development
Enquire if the software is done under the guidance of doctors
"Leave a review" modalwindow within the software
User may say that they will refer to others as well.
Customer waits for the result
Completes the review section of the software
Will see the corrections of others who have completed the test
Goals & motivations
At each step, what is a person's primary goal or motivation? ("Help me..." or "Help me avoid...")
Help me in the usage of the software
Help me to check the correctness of the software
Help me to feed data into the software
Help me to reach this CKD prediction section
Help me to correctly enter the data into the software
Help me in various payment modes that can be done
Help me to avoid fear of the various parameters
Help me feel confident about the software
Help me feel good about my decision to go on this checkup.
Help me to cooperate with the nurse
Help me to get the result as no CKD
Help me to realize my food habits
Help me to know that exercise is the best remedy
Positive moments
What steps does a typical person find enjoyable, productive, fun, motivating, delightful, or exciting?
It's productive withoutwasting doctor's time
I have never taken so many tests before
Have to read reviews about others experience
Excitement about the result in few seconds
Payment is cheap and results quick
Persons get excited about the simplicity of the process
Patients are happy about the staff treatment
People leave the CKD system with satisfaction
Happy to know the result if it is Negative
Negative moments
What steps does a typical person find frustrating, confusing, angering, costly, or time-consuming?
People get frustrated if they don't get doctor appointment
People feel that CKD is asking for doing too many tests
CKD may be time consuming
Whether the cost paidfor the test is really worth it
People expressed difficulty in the availability of the CKD system
People find that the software is asking too many questions
People feel that the process may be incomplete
People feel that the result should be cross checked with the doctor
Need to check reviews for the usage of the CKD software
Areas of opportunity
How might we make each step better? What ideas do we have? What have others suggested?
Software should be reliable
It should be self explanatory
Software ui
Should provide software review stars
CKD should be visible and comfortable
Payment should be less
People feel it should be operated by doctors
CKD should be an expert system to assist doctors
Patients' data should be kept confidential
Doctor's reference name should be linked with patient

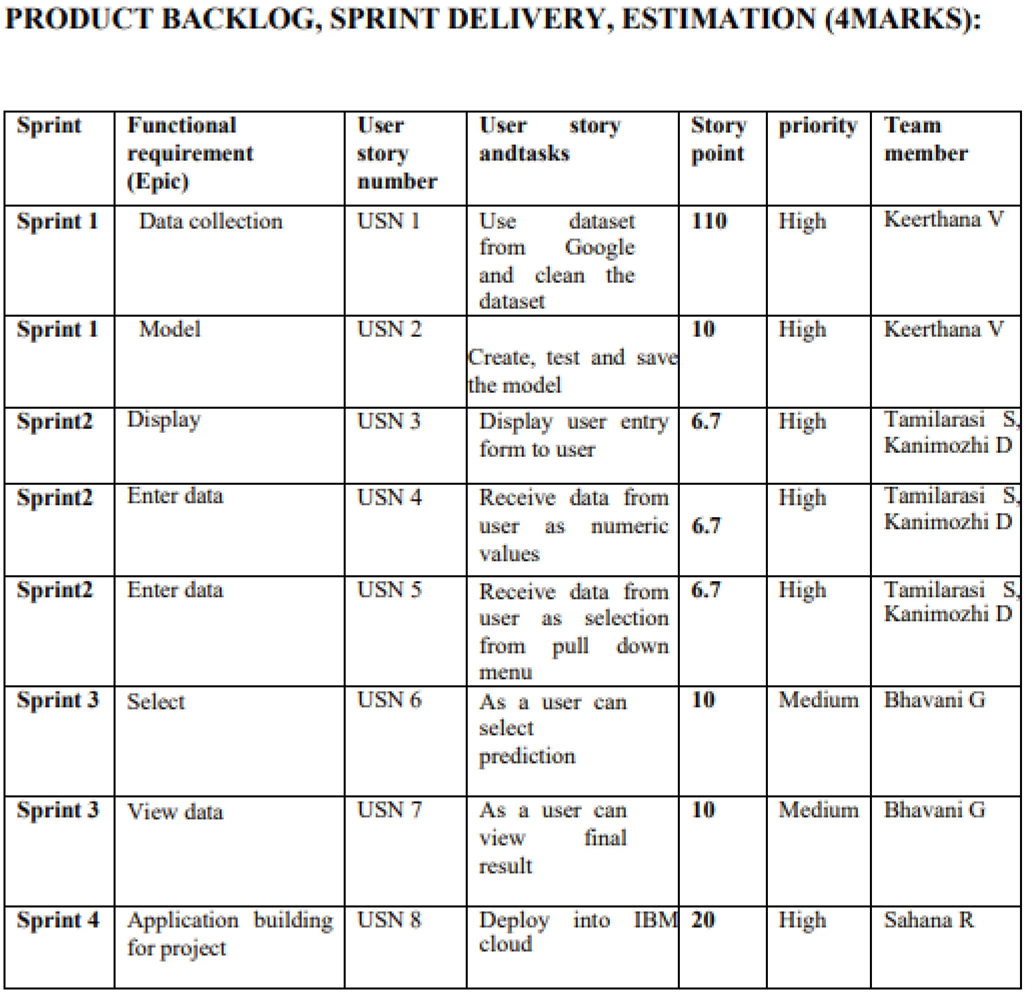
Fig 5.2.2 Customer Journey Map

6. PROJECT PLANNING & SCHEDULING

During the Project Planning Phase we have done Project planning template, Milestone and activity list and Jira sprint delivery plan.

6.1 Sprint Planning & Estimation:

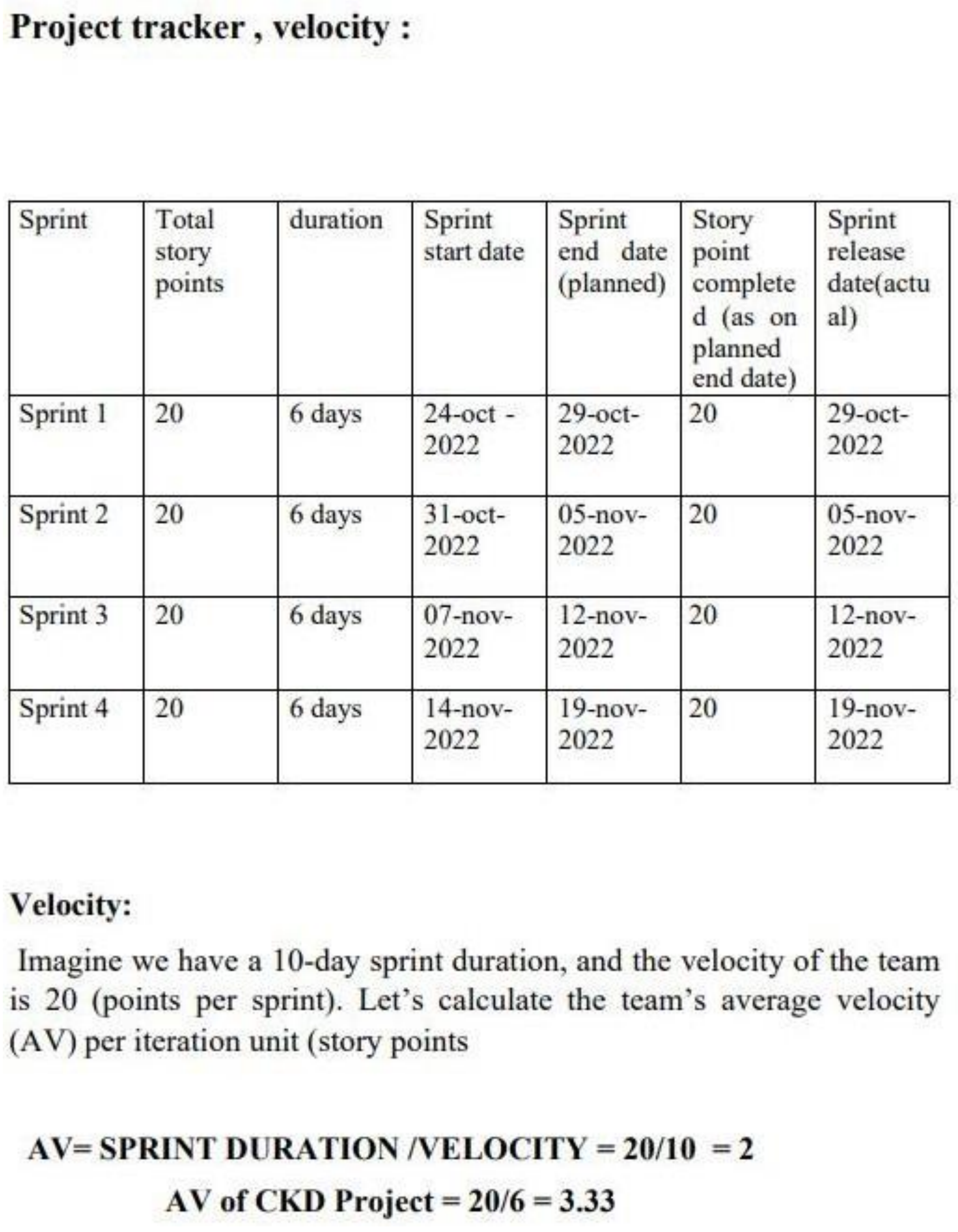
A project plan template is a document that creates a standard format for a project plan. Typically, it contains a list of the essential elements of a project, such as stakeholders, scope, timelines, estimated cost and communication methods. The project manager typically lists the information based on the assignment.



PRODUCT BACKLOG, SPRINT DELIVERY, ESTIMATION (4MARKS):

Sprint	Functional requirement (Epic)	User story number	User story andtasks	Story point	priority	Team member
Sprint 1	Data collection	USN 1	Use dataset from Google and clean the dataset	110	High	Keerthana V
Sprint 1	Model	USN 2	Create, test and save the model	10	High	Keerthana V
Sprint2	Display	USN 3	Display user entry form to user	6.7	High	Tamilarasi S, Kanimozhi D
Sprint2	Enter data	USN 4	Receive data from user as numeric values	6.7	High	Tamilarasi S, Kanimozhi D
Sprint2	Enter data	USN 5	Receive data from user as selection from pull down menu	6.7	High	Tamilarasi S, Kanimozhi D
Sprint 3	Select	USN 6	As a user can select prediction	10	Medium	Bhavani G
Sprint 3	View data	USN 7	As a user can view final result	10	Medium	Bhavani G
Sprint 4	Application building for project	USN 8	Deploy into IBM cloud	20	High	Sahana R

Table 6.1.1 Sprint Planning & Estimation



Project tracker , velocity :

Sprint	Total story points	duration	Sprint start date	Sprint end date (planned)	Story point completed (as on planned end date)	Sprint release date(actual)
Sprint 1	20	6 days	24-oct - 2022	29-oct-2022	20	29-oct-2022
Sprint 2	20	6 days	31-oct-2022	05-nov-2022	20	05-nov-2022
Sprint 3	20	6 days	07-nov-2022	12-nov-2022	20	12-nov-2022
Sprint 4	20	6 days	14-nov-2022	19-nov-2022	20	19-nov-2022

Velocity:

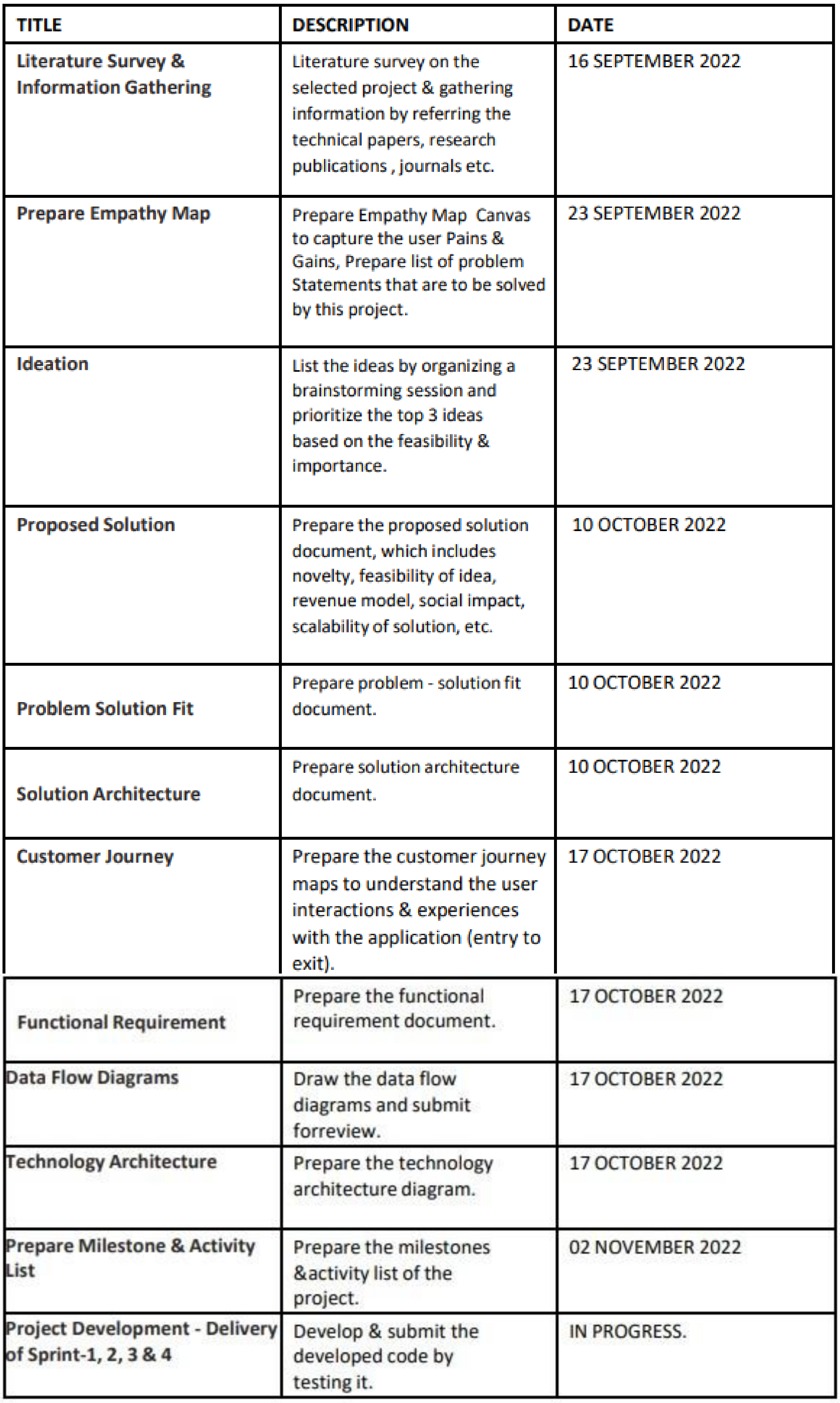
Imagine we have a 10-day sprint duration, and the velocity of the team is 20 (points per sprint). Let's calculate the team's average velocity (AV) per iteration unit (story points

AV= SPRINT DURATION /VELOCITY = 20/10 = 2

AV of CKD Project = 20/6 = 3.33

6.2 SPRINT DELIVERY SCHEDULE:

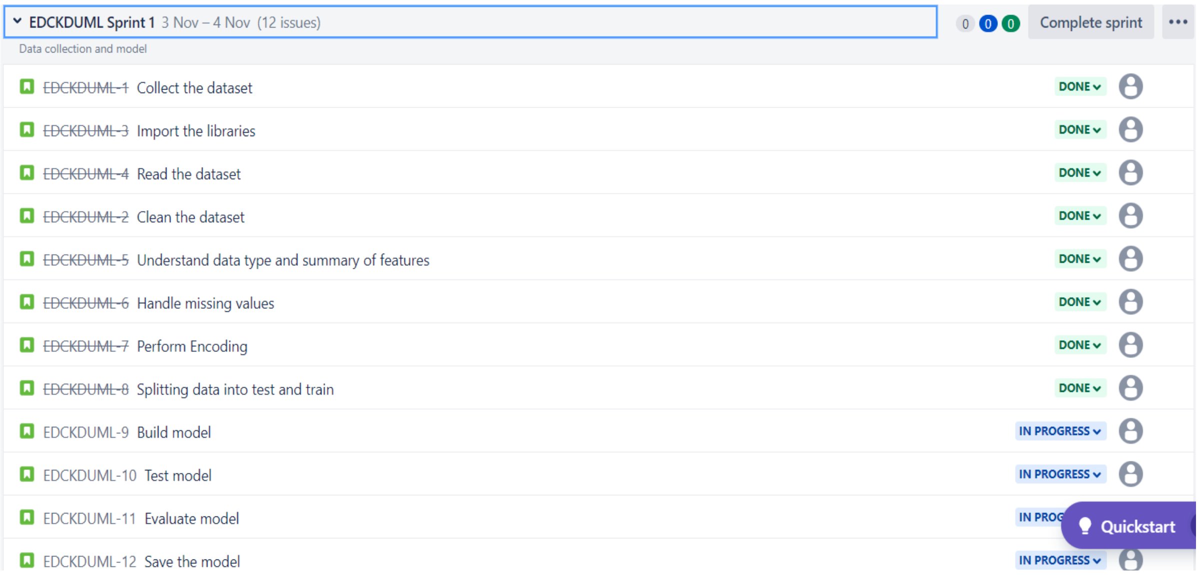
A milestone list is a project management document that identifies all project milestones. A milestone is a significant event or a point in a project. It represents nothing more than a moment in time; hence, when scheduling, milestones should be assigned zero duration.



TITLE	DESCRIPTION	DATE
Literature Survey & Information Gathering	Literature survey on the selected project & gathering information by referring the technical papers, research publications , journals etc.	16 SEPTEMBER 2022
Prepare Empathy Map	Prepare Empathy Map Canvas to capture the user Pains & Gains, Prepare list of problem Statements that are to be solved by this project.	23 SEPTEMBER 2022
Ideation	List the ideas by organizing a brainstorming session and prioritize the top 3 ideas based on the feasibility & importance.	23 SEPTEMBER 2022
Proposed Solution	Prepare the proposed solution document, which includes novelty, feasibility of idea, revenue model, social impact, scalability of solution, etc.	10 OCTOBER 2022
Problem Solution Fit	Prepare problem - solution fit document.	10 OCTOBER 2022
Solution Architecture	Prepare solution architecture document.	10 OCTOBER 2022
Customer Journey	Prepare the customer journey maps to understand the user interactions & experiences with the application (entry to exit).	17 OCTOBER 2022
Functional Requirement	Prepare the functional requirement document.	17 OCTOBER 2022
Data Flow Diagrams	Draw the data flow diagrams and submit forreview.	17 OCTOBER 2022
Technology Architecture	Prepare the technology architecture diagram.	17 OCTOBER 2022
Prepare Milestone & Activity List	Prepare the milestones &activity list of the project.	02 NOVEMBER 2022
Project Development - Delivery of Sprint-1, 2, 3 & 4	Develop & submit the developed code by testing it.	IN PROGRESS.

6.3 REPORI’S ÏROM JIRA:

SPRINT 1:



EDCKDUML Sprint 1 3 Nov – 4 Nov (12 issues)

0 0 0 Complete sprint •••

Data collection and model

~~EDCKDUML-1~~ Collect the dataset DONE

~~EDCKDUML-3~~ Import the libraries DONE

~~EDCKDUML-4~~ Read the dataset DONE

~~EDCKDUML-2~~ Clean the dataset DONE

~~EDCKDUML-5~~ Understand data type and summary of features DONE

~~EDCKDUML-6~~ Handle missing values DONE

~~EDCKDUML-7~~ Perform Encoding DONE

~~EDCKDUML-8~~ Splitting data into test and train DONE

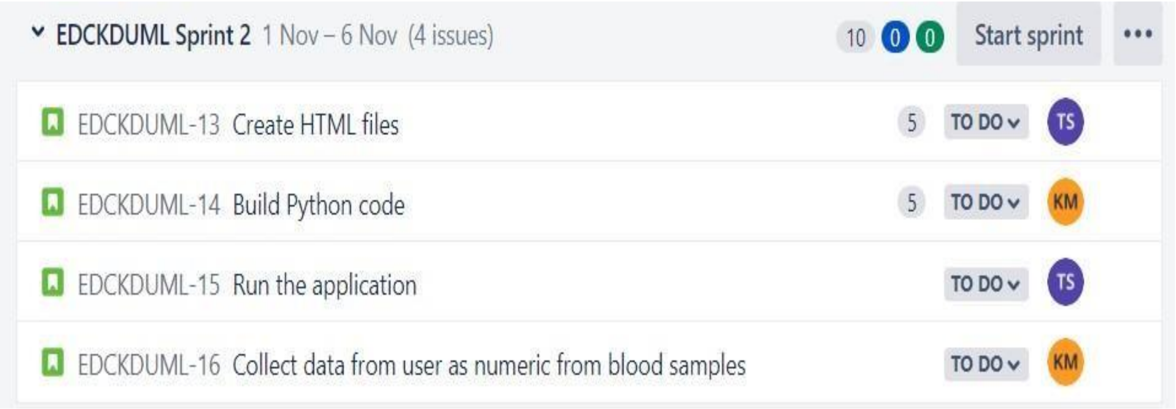
EDCKDUML-9 Build model IN PROGRESS

EDCKDUML-10 Test model IN PROGRESS

EDCKDUML-11 Evaluate model IN PROG

EDCKDUML-12 Save the model IN PROGRESS

SPRINT 2:



EDCKDUML Sprint 2 1 Nov – 6 Nov (4 issues)

10 0 0 Start sprint •••

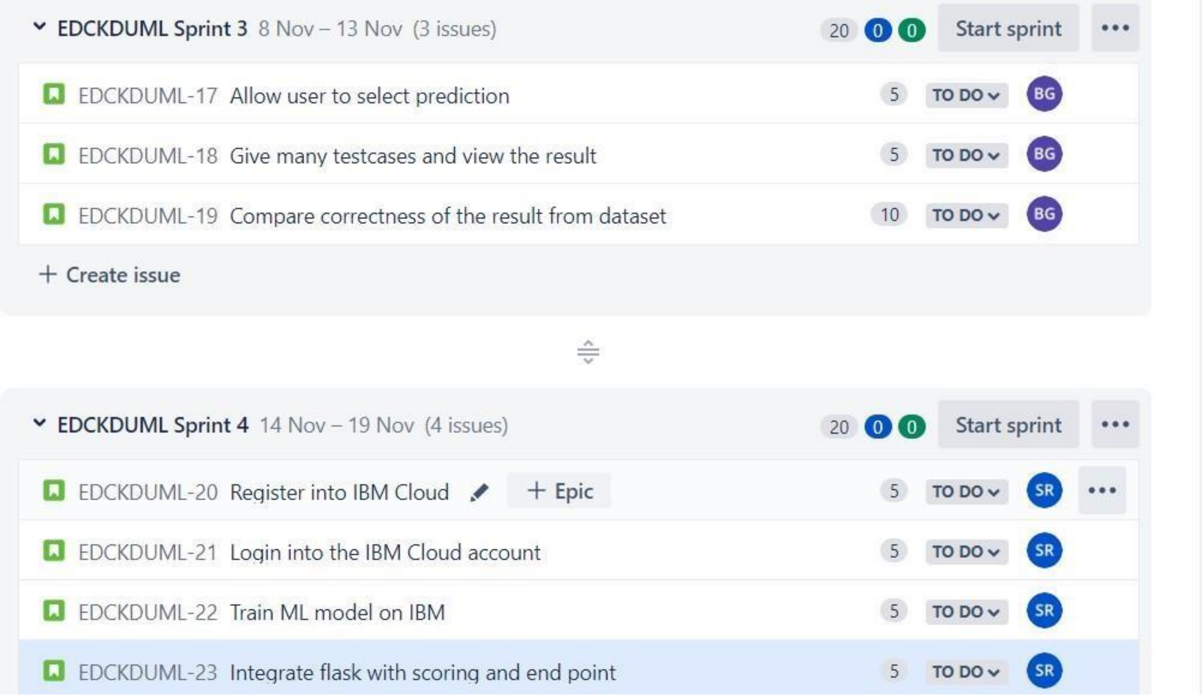
EDCKDUML-13 Create HTML files 5 TO DO TS

EDCKDUML-14 Build Python code 5 TO DO KM

EDCKDUML-15 Run the application TO DO TS

EDCKDUML-16 Collect data from user as numeric from blood samples TO DO KM

SPRINT 3 and 4:



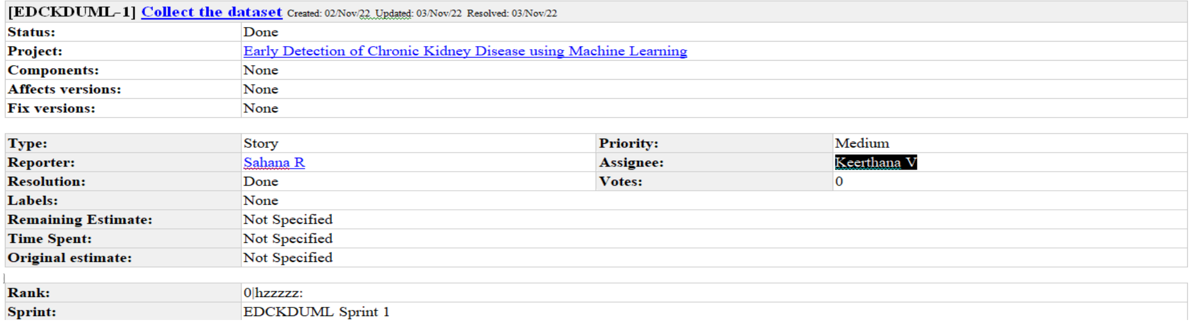
EDCKDUML Sprint 3 8 Nov – 13 Nov (3 issues) 20 0 0 Start sprint •••

EDCKDUML-17	Allow user to select prediction	5	TO DO	BG
EDCKDUML-18	Give many testcases and view the result	5	TO DO	BG
EDCKDUML-19	Compare correctness of the result from dataset	10	TO DO	BG

EDCKDUML Sprint 4 14 Nov – 19 Nov (4 issues) 20 0 0 Start sprint •••

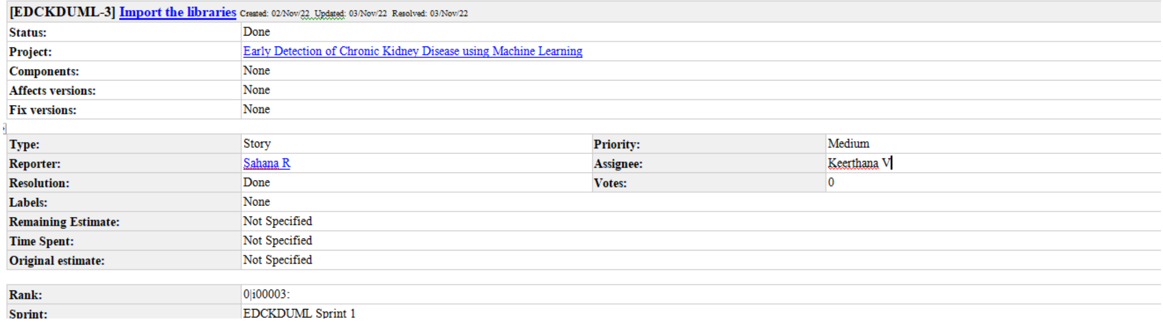
EDCKDUML-20	Register into IBM Cloud	5	TO DO	SR
EDCKDUML-21	Login into the IBM Cloud account	5	TO DO	SR
EDCKDUML-22	Train ML model on IBM	5	TO DO	SR
EDCKDUML-23	Integrate flask with scoring and end point	5	TO DO	SR

SCREENSHOTS:



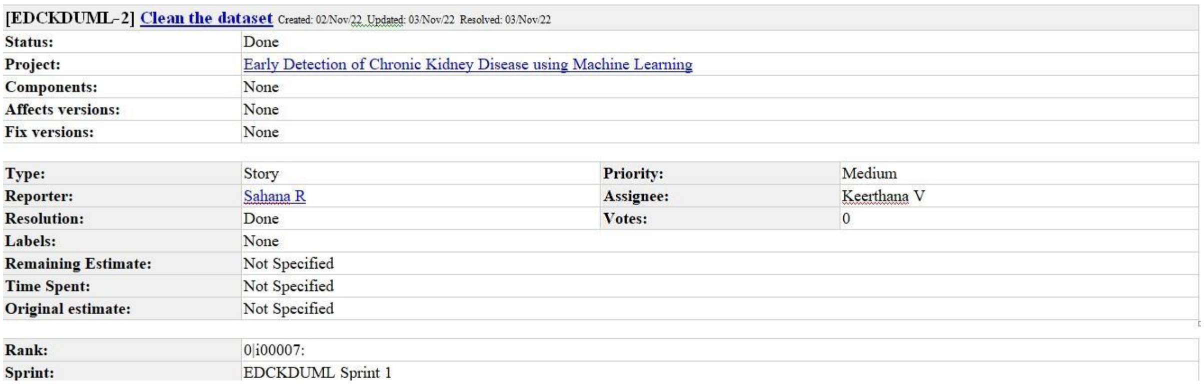
[EDCKDUML-1] Collect the dataset Created: 02/Nov/22 Updated: 03/Nov/22 Resolved: 03/Nov/22

Status:	Done		
Project:	Early Detection of Chronic Kidney Disease using Machine Learning		
Components:	None		
Affects versions:	None		
Fix versions:	None		
Type:	Story	**Priority:**	Medium
Reporter:	Sahana R	**Assignee:**	Keerthana V
Resolution:	Done	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		
Rank:	0|hzzzzz:		
Sprint:	EDCKDUML Sprint 1		



[EDCKDUML-3] Import the libraries Created: 02/Nov/22 Updated: 03/Nov/22 Resolved: 03/Nov/22

Status:	Done		
Project:	Early Detection of Chronic Kidney Disease using Machine Learning		
Components:	None		
Affects versions:	None		
Fix versions:	None		
Type:	Story	**Priority:**	Medium
Reporter:	Sahana R	**Assignee:**	Keerthana V
Resolution:	Done	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		
Rank:	0|i00003:		
Sprint:	EDCKDUML Sprint 1		

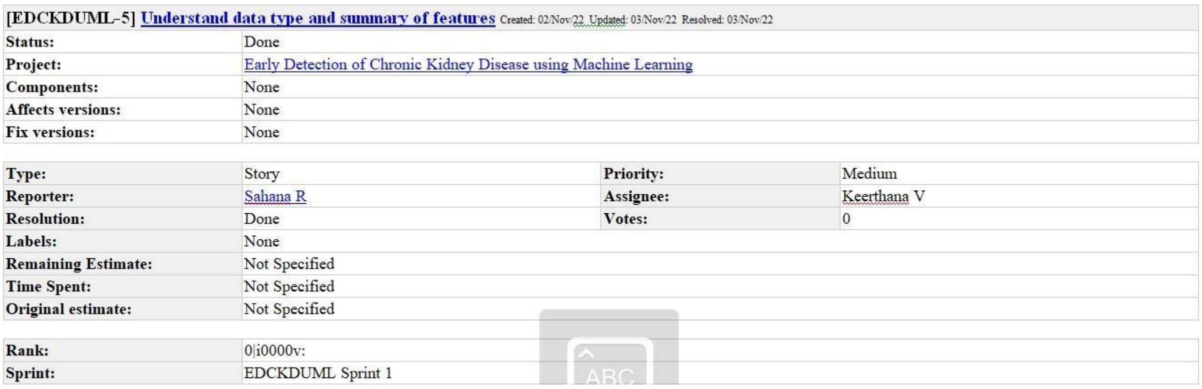


[EDCKDUML-2] Clean the dataset Created: 02/Nov/22 Updated: 03/Nov/22 Resolved: 03/Nov/22

Status:	Done
Project:	Early Detection of Chronic Kidney Disease using Machine Learning
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Sahana R	**Assignee:**	Keerthana V
Resolution:	Done	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		

Rank:	0|i00007:
Sprint:	EDCKDUML Sprint 1

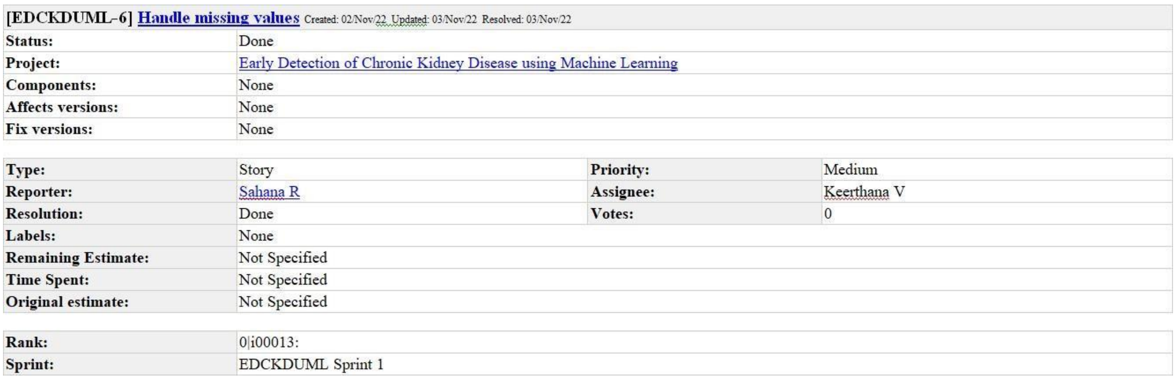


[EDCKDUML-5] Understand data type and summary of features Created: 02/Nov/22 Updated: 03/Nov/22 Resolved: 03/Nov/22

Status:	Done
Project:	Early Detection of Chronic Kidney Disease using Machine Learning
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Sahana R	**Assignee:**	Keerthana V
Resolution:	Done	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		

Rank:	0|i0000v:
Sprint:	EDCKDUML Sprint 1

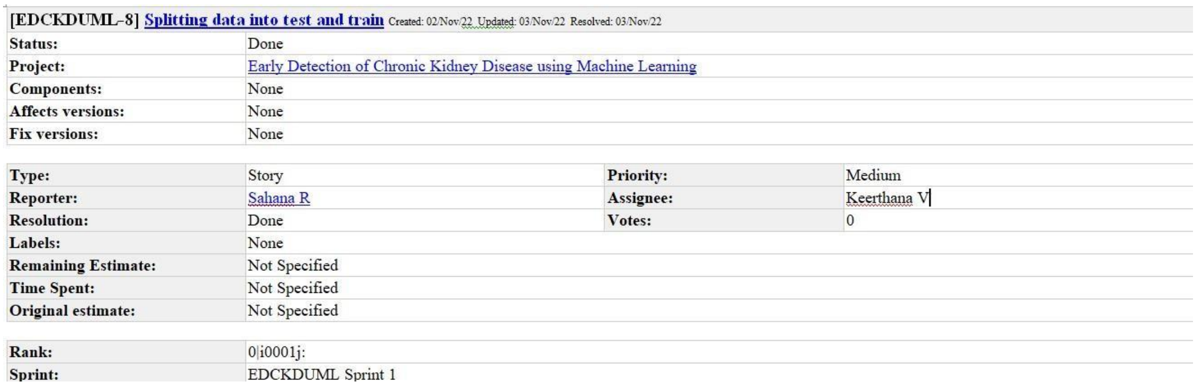


[EDCKDUML-6] Handle missing values Created: 02/Nov/22 Updated: 03/Nov/22 Resolved: 03/Nov/22

Status:	Done
Project:	Early Detection of Chronic Kidney Disease using Machine Learning
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Sahana R	**Assignee:**	Keerthana V
Resolution:	Done	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		

Rank:	0|i00013:
Sprint:	EDCKDUML Sprint 1

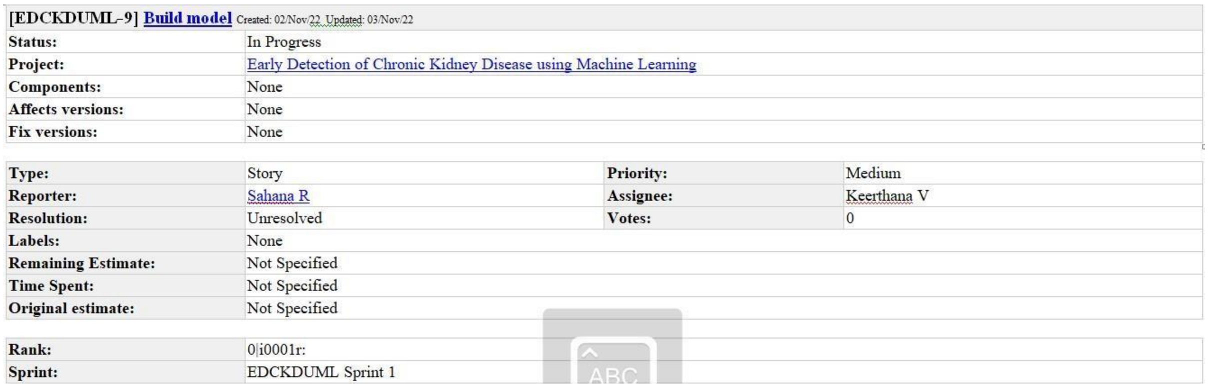


[EDCKDUML-8] Splitting data into test and train Created: 02/Nov/22 Updated: 03/Nov/22 Resolved: 03/Nov/22

Status:	Done
Project:	Early Detection of Chronic Kidney Disease using Machine Learning
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Sahana R	**Assignee:**	Keerthana V
Resolution:	Done	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		

Rank:	0|i0001j:
Sprint:	EDCKDUML Sprint 1

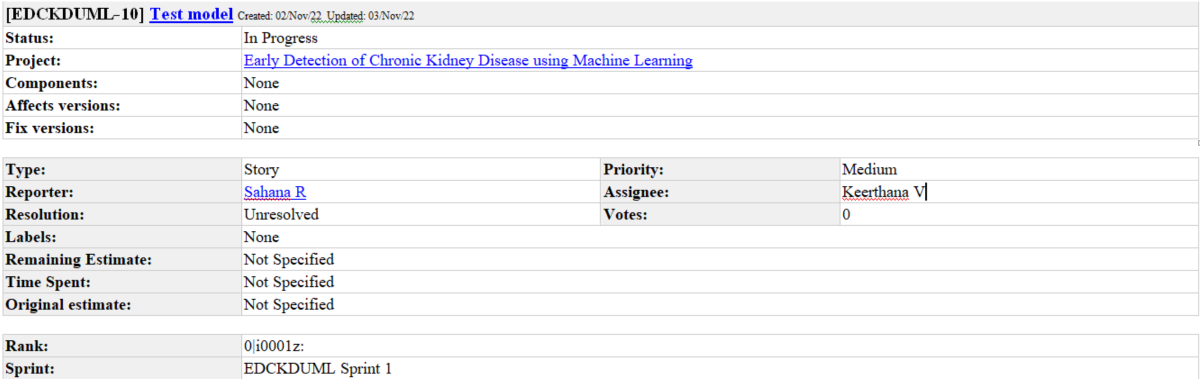


[EDCKDUML-9] Build model Created: 02/Nov/22 Updated: 03/Nov/22

Status:	In Progress
Project:	Early Detection of Chronic Kidney Disease using Machine Learning
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Sahana R	**Assignee:**	Keerthana V
Resolution:	Unresolved	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		

Rank:	0|i0001r:
Sprint:	EDCKDUML Sprint 1

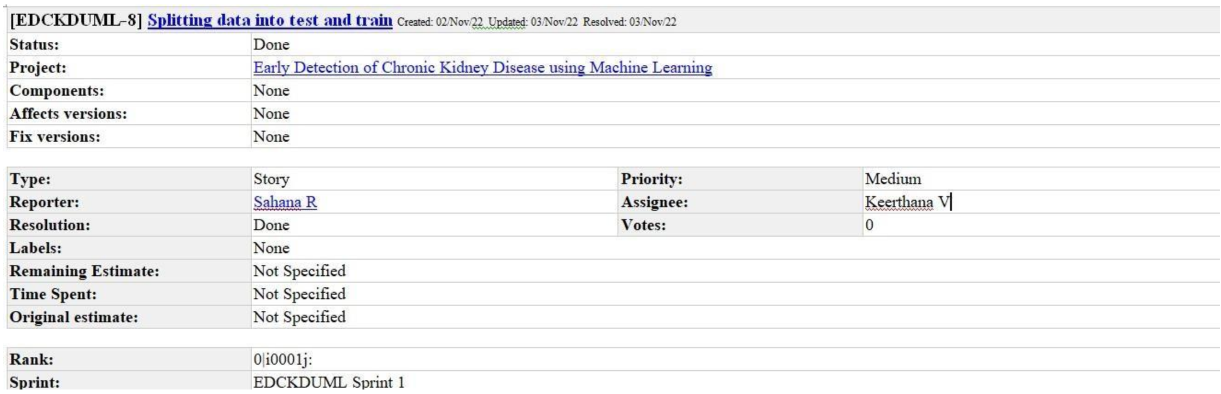


[EDCKDUML-10] Test model Created: 02/Nov/22 Updated: 03/Nov/22

Status:	In Progress
Project:	Early Detection of Chronic Kidney Disease using Machine Learning
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Sahana R	**Assignee:**	Keerthana V
Resolution:	Unresolved	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		

Rank:	0|i0001z:
Sprint:	EDCKDUML Sprint 1

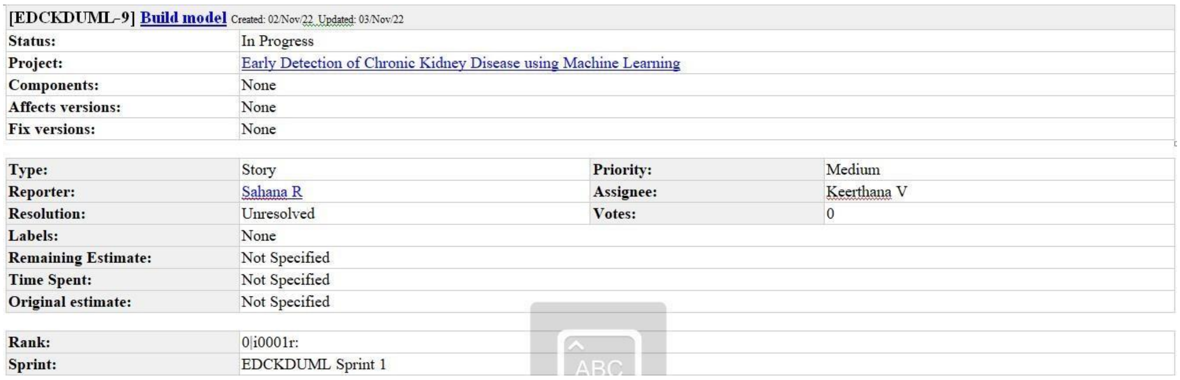


[EDCKDUML-8] Splitting data into test and train Created: 02/Nov/22 Updated: 03/Nov/22 Resolved: 03/Nov/22

Status:	Done
Project:	Early Detection of Chronic Kidney Disease using Machine Learning
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Sahana R	**Assignee:**	Keerthana V
Resolution:	Done	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		

Rank:	0|i0001j:
Sprint:	EDCKDUML Sprint 1

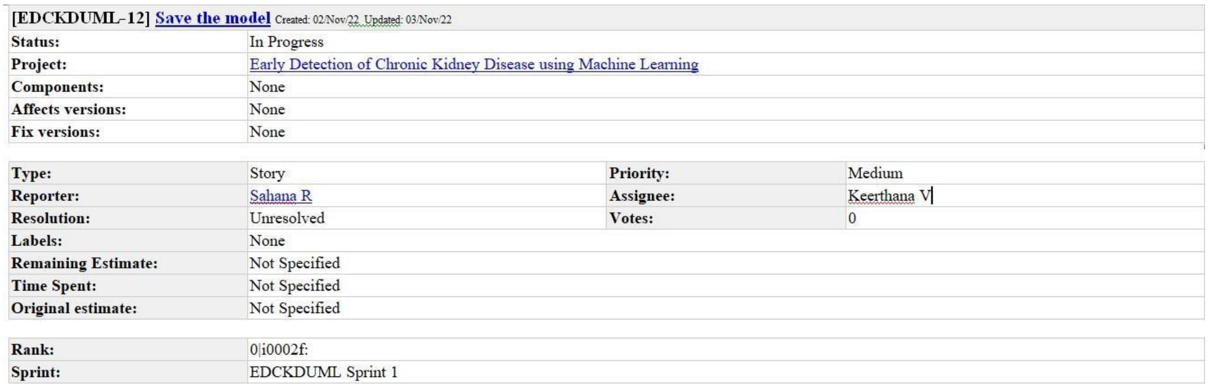


[EDCKDUML-9] Build model Created: 02/Nov/22 Updated: 03/Nov/22

Status:	In Progress
Project:	Early Detection of Chronic Kidney Disease using Machine Learning
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Sahana R	**Assignee:**	Keerthana V
Resolution:	Unresolved	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		

Rank:	0|i0001r:
Sprint:	EDCKDUML Sprint 1

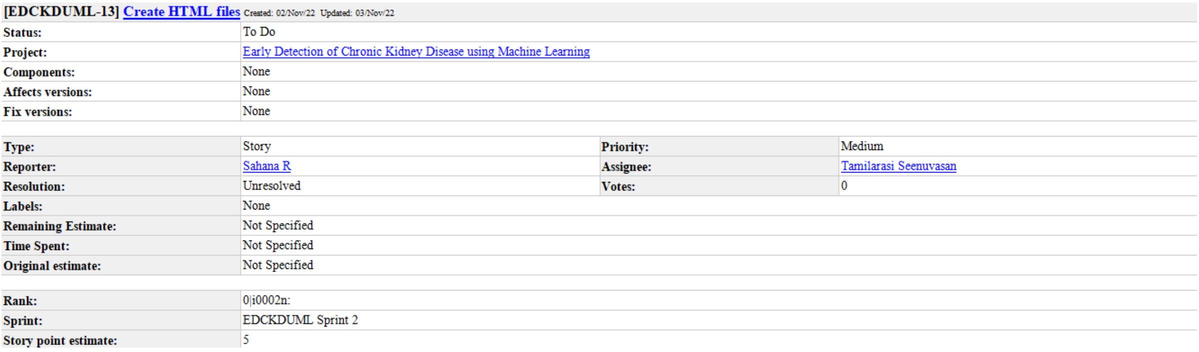


[EDCKDUML-12] Save the model Created: 02/Nov/22 Updated: 03/Nov/22

Status:	In Progress
Project:	Early Detection of Chronic Kidney Disease using Machine Learning
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Sahana R	**Assignee:**	Keerthana V
Resolution:	Unresolved	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		

Rank:	0|i0002f:
Sprint:	EDCKDUML Sprint 1

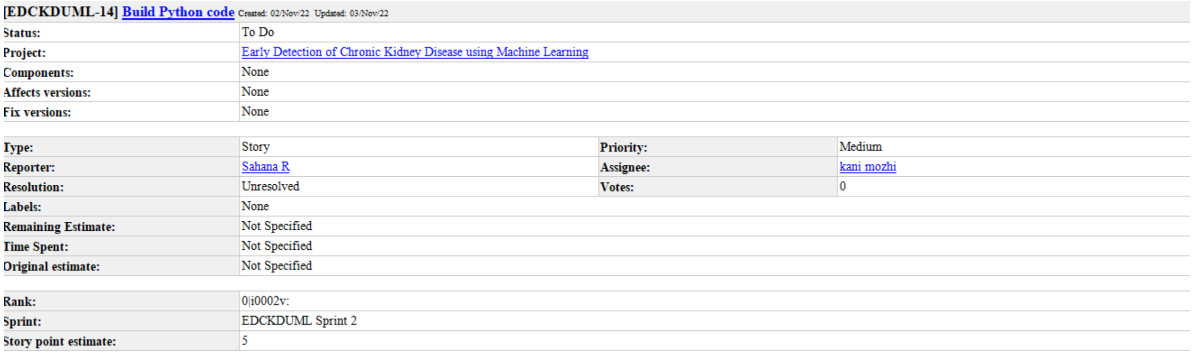


[EDCKDUML-13] Create HTML files Created: 02/Nov/22 Updated: 03/Nov/22

Status:	To Do
Project:	Early Detection of Chronic Kidney Disease using Machine Learning
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Sahana R	**Assignee:**	Tamilarasi Seenuvasan
Resolution:	Unresolved	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		

Rank:	0|i0002n:
Sprint:	EDCKDUML Sprint 2
Story point estimate:	5

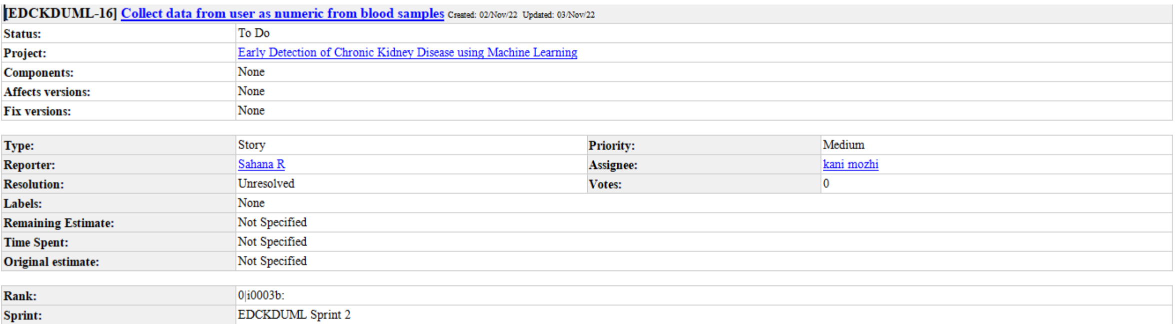


[EDCKDUML-14] Build Python code Created: 02/Nov/22 Updated: 03/Nov/22

Status:	To Do
Project:	Early Detection of Chronic Kidney Disease using Machine Learning
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Sahana R	**Assignee:**	kani mozhi
Resolution:	Unresolved	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		

Rank:	0|i0002v:
Sprint:	EDCKDUML Sprint 2
Story point estimate:	5

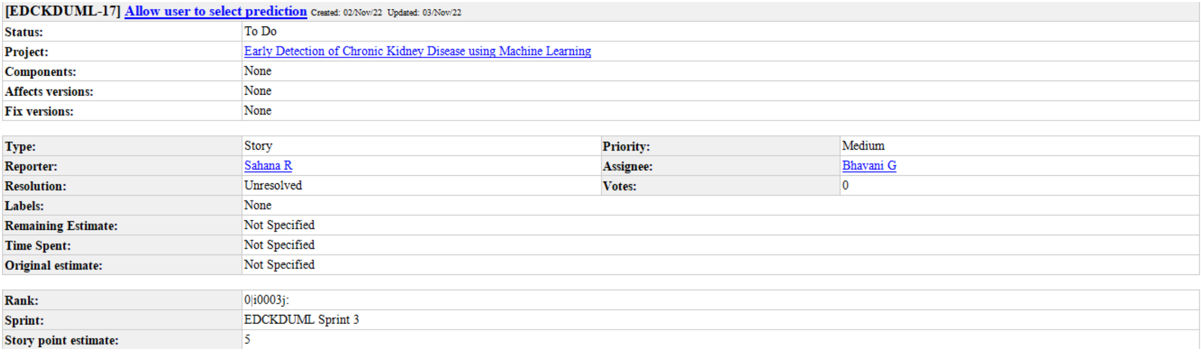


[EDCKDUML-16] Collect data from user as numeric from blood samples Created: 02/Nov/22 Updated: 03/Nov/22

Status:	To Do
Project:	Early Detection of Chronic Kidney Disease using Machine Learning
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Sahana R	**Assignee:**	kani mozhi
Resolution:	Unresolved	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		

Rank:	0|i0003b:
Sprint:	EDCKDUML Sprint 2

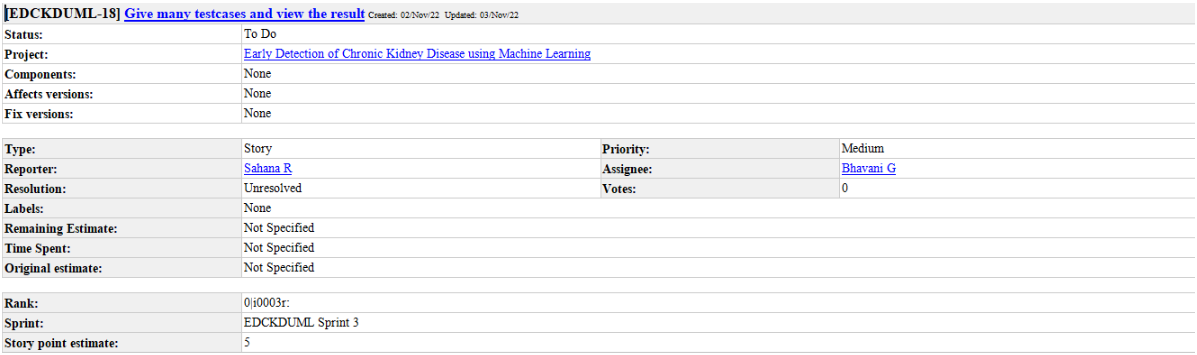


[EDCKDUML-17] Allow user to select prediction Created: 02/Nov/22 Updated: 03/Nov/22

Status:	To Do
Project:	Early Detection of Chronic Kidney Disease using Machine Learning
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Sahana R	**Assignee:**	Bhavani G
Resolution:	Unresolved	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		

Rank:	0|i0003j:
Sprint:	EDCKDUML Sprint 3
Story point estimate:	5

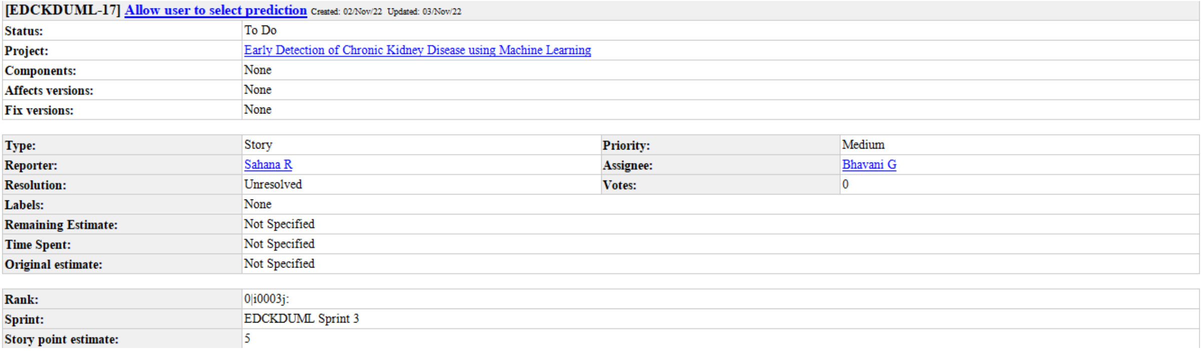


[EDCKDUML-18] Give many testcases and view the result Created: 02/Nov/22 Updated: 03/Nov/22

Status:	To Do
Project:	Early Detection of Chronic Kidney Disease using Machine Learning
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Sahana R	**Assignee:**	Bhavani G
Resolution:	Unresolved	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		

Rank:	0|i0003r:
Sprint:	EDCKDUML Sprint 3
Story point estimate:	5

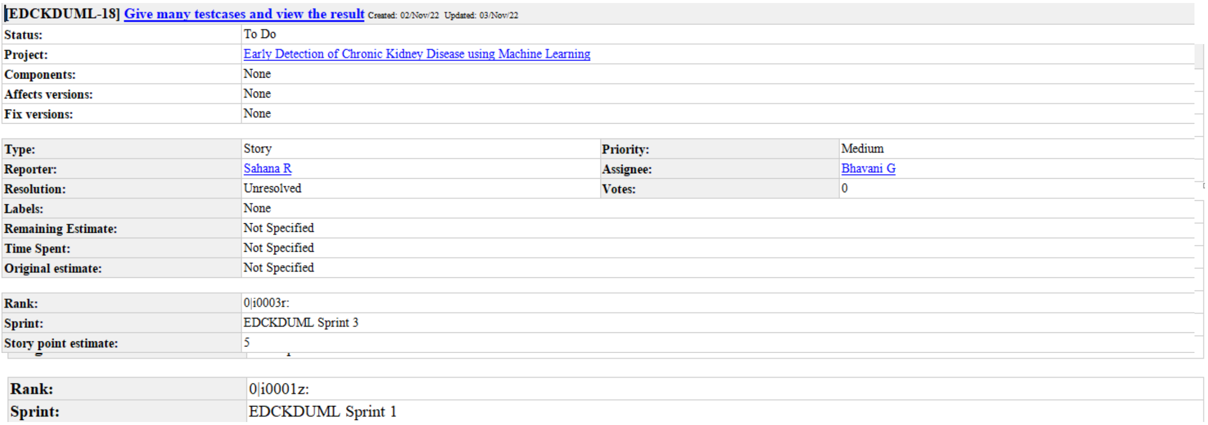


[EDCKDUML-17] Allow user to select prediction Created: 02/Nov/22 Updated: 03/Nov/22

Status:	To Do
Project:	Early Detection of Chronic Kidney Disease using Machine Learning
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Sahana R	**Assignee:**	Bhavani G
Resolution:	Unresolved	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		

Rank:	0|i0003j:
Sprint:	EDCKDUML Sprint 3
Story point estimate:	5



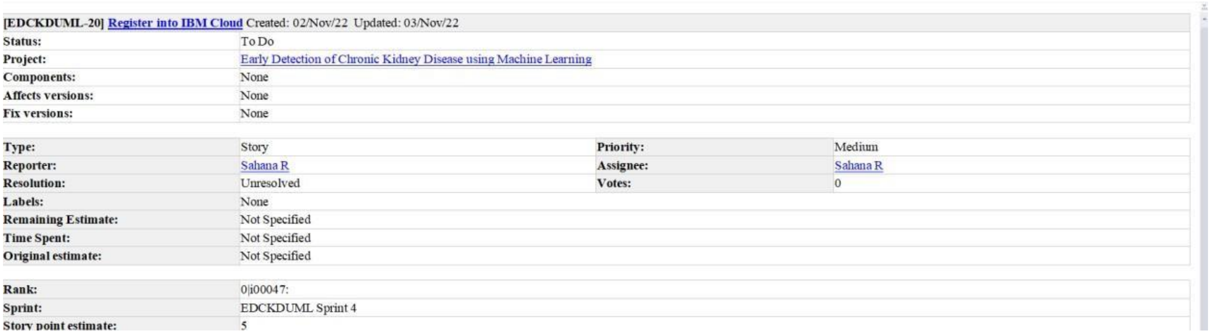
[EDCKDUML-18] Give many testcases and view the result Created: 02/Nov/22 Updated: 03/Nov/22

Status:	To Do
Project:	Early Detection of Chronic Kidney Disease using Machine Learning
Components:	None
Affects versions:	None
Fix versions:	None

Type:	Story	**Priority:**	Medium
Reporter:	Sahana R	**Assignee:**	Bhavani G
Resolution:	Unresolved	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		

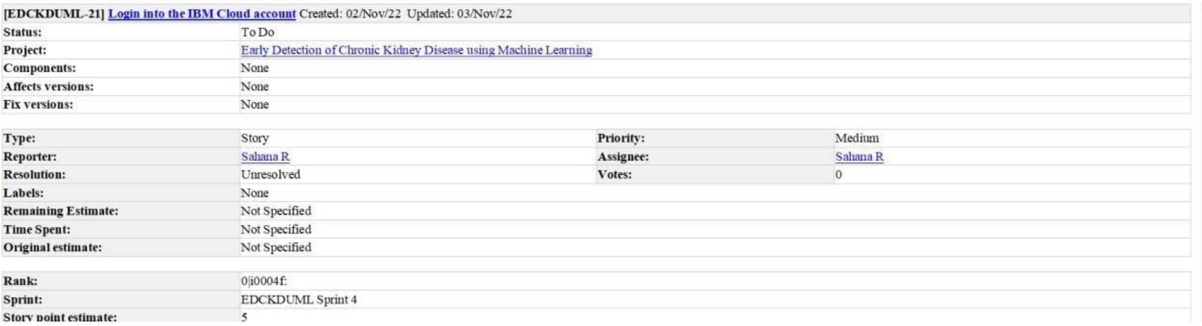
Rank:	0|i0003r:
Sprint:	EDCKDUML Sprint 3
Story point estimate:	5

Rank:	0|i0001z:
Sprint:	EDCKDUML Sprint 1



[EDCKDUML-20] Register into IBM Cloud Created: 02/Nov/22 Updated: 03/Nov/22

Status:	To Do		
Project:	Early Detection of Chronic Kidney Disease using Machine Learning		
Components:	None		
Affects versions:	None		
Fix versions:	None		
Type:	Story	**Priority:**	Medium
Reporter:	Sahana R	**Assignee:**	Sahana R
Resolution:	Unresolved	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		
Rank:	0|i00047:		
Sprint:	EDCKDUML Sprint 4		
Story point estimate:	5		



[EDCKDUML-21] Login into the IBM Cloud account Created: 02/Nov/22 Updated: 03/Nov/22

Status:	To Do		
Project:	Early Detection of Chronic Kidney Disease using Machine Learning		
Components:	None		
Affects versions:	None		
Fix versions:	None		
Type:	Story	**Priority:**	Medium
Reporter:	Sahana R	**Assignee:**	Sahana R
Resolution:	Unresolved	**Votes:**	0
Labels:	None		
Remaining Estimate:	Not Specified		
Time Spent:	Not Specified		
Original estimate:	Not Specified		
Rank:	0|i0004f:		
Sprint:	EDCKDUML Sprint 4		
Story point estimate:	5		

7.0 CODING AND SOLUTIONING:

During the Project Development Phase we have done four Sprints they are Sprint 1, Sprint 2, Sprint 3 and Sprint 4.In Agile product development, a sprint is a set period of time during which specific work has to be completed and made ready for review.

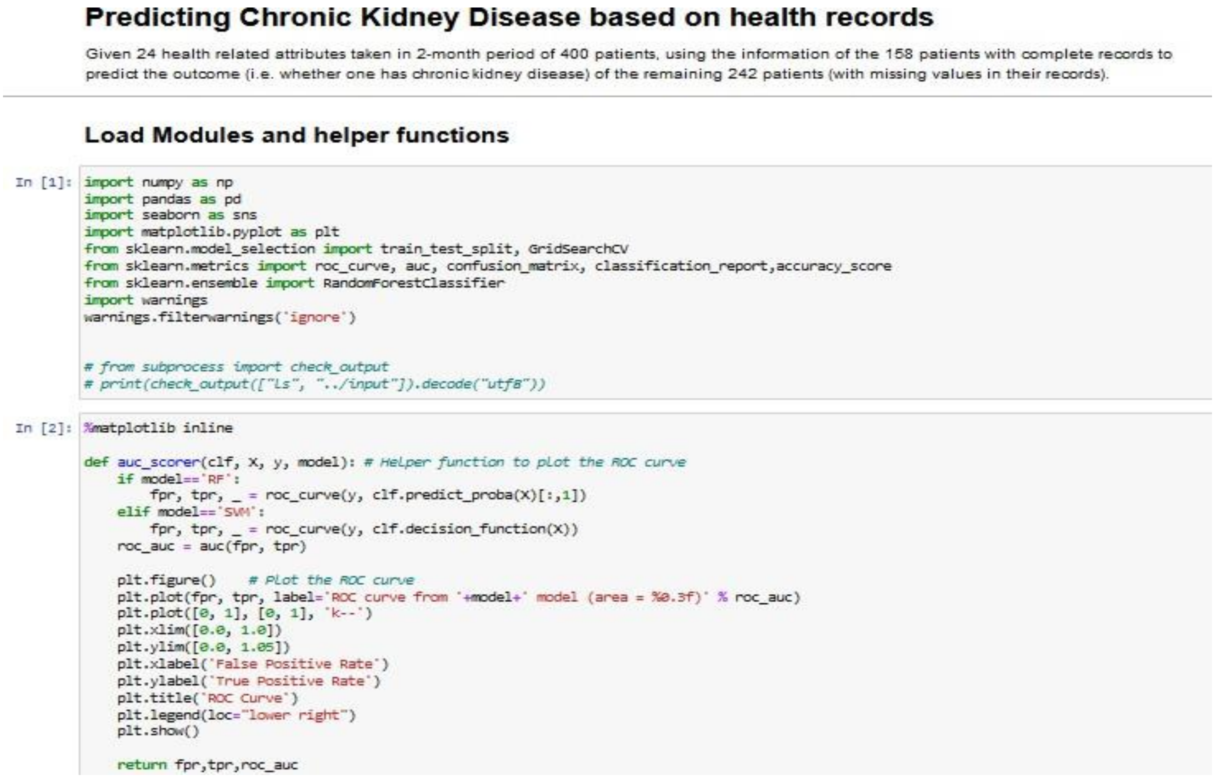
Each sprint begins with a planning meeting. During the meeting, the product owner (the person requesting the work) and the development team agree upon exactly what work will be accomplished during the sprint. The development team has the final say when it comes to determining how much work can realistically be accomplished during the sprint, and the product owner has the final say on what criteria need to be met for the work to be approved and accepted.

The duration of a sprint is determined by the scrum master, the team's facilitator and manager of the Scrum framework. Once the team reaches a consensus for how many days a sprint should last, all future sprints should be the same. Traditionally, a sprint lasts 30 days.

After a sprint begins, the product owner must step back and let the team do their work. During the sprint, the team holds daily stand-up meetings to discuss progress and brainstorm solutions to challenges. The project owner may attend these meetings as an observer but is not allowed to participate unless it is to answer question. The project owner may not make requests for changes during a sprint and only the scrum master or project manager has the power to interrupt or stop the sprint.

At the end of the sprint, the team presents its completed work to the project owner and the project owner uses the criteria established at the sprint planning meeting to either accept or reject the work.

7.1 Feature 1:



Predicting Chronic Kidney Disease based on health records

Given 24 health related attributes taken in 2-month period of 400 patients, using the information of the 158 patients with complete records to predict the outcome (i.e. whether one has chronic kidney disease) of the remaining 242 patients (with missing values in their records).

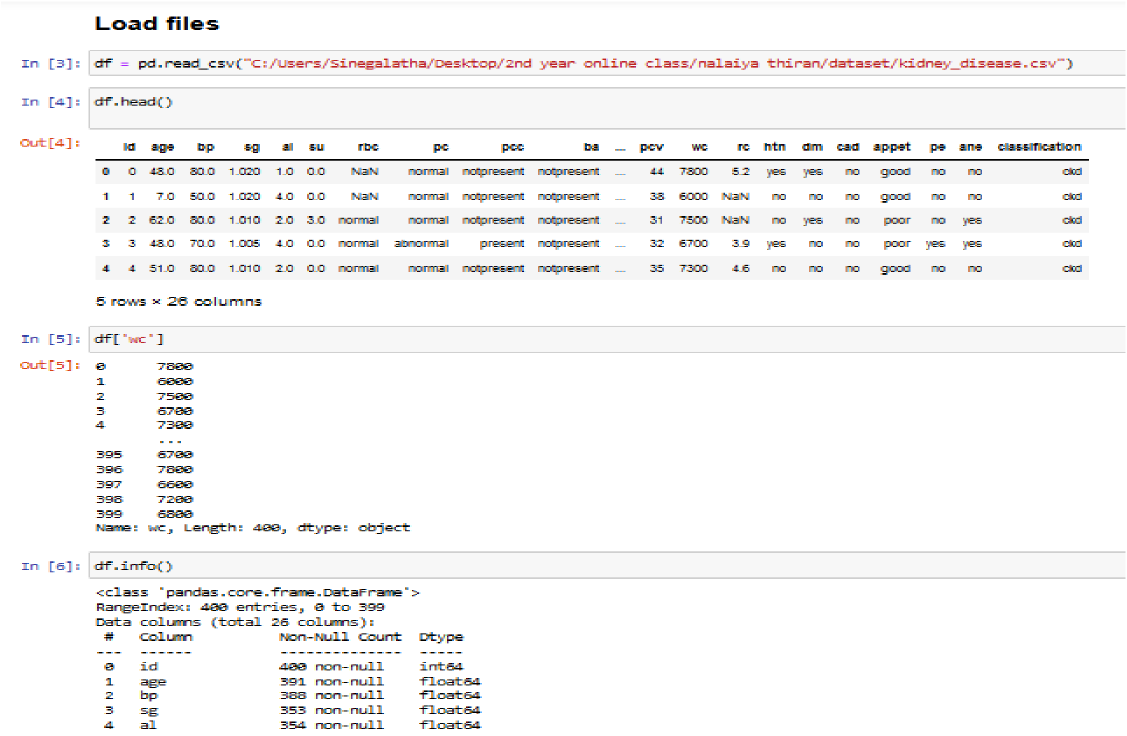
Load Modules and helper functions

In [1]:

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.metrics import roc_curve, auc, confusion_matrix, classification_report,accuracy_score
from sklearn.ensemble import RandomForestClassifier
import warnings
warnings.filterwarnings('ignore')


# from subprocess import check_output
# print(check_output(["ls", "../input"]).decode("utf8"))
```

In [2]:

```
%matplotlib inline

def auc_scorer(clf, X, y, model): # Helper function to plot the ROC curve
    if model=='RF':
        fpr, tpr, _ = roc_curve(y, clf.predict_proba(X)[:,1])
    elif model=='SVM':
        fpr, tpr, _ = roc_curve(y, clf.decision_function(X))
    roc_auc = auc(fpr, tpr)

    plt.figure()    # Plot the ROC curve
    plt.plot(fpr, tpr, label='ROC curve from '+model+' model (area = %0.3f)' % roc_auc)
    plt.plot([0, 1], [0, 1], 'k--')
    plt.xlim([0.0, 1.0])
    plt.ylim([0.0, 1.05])
    plt.xlabel('False Positive Rate')
    plt.ylabel('True Positive Rate')
    plt.title('ROC Curve')
    plt.legend(loc="lower right")
    plt.show()

    return fpr,tpr,roc_auc
```



Load files

In [3]:

```
df = pd.read_csv("C:/Users/Sinegalatha/Desktop/2nd year online class/nalaiya thiran/dataset/kidney_disease.csv")
```

In [4]:

```
df.head()
```

Out[4]:

	id	age	bp	sg	al	su	rbc	pc	pcc	ba	...	pcv	wc	rc	htn	dm	cad	appet	pe	ane	classification
0	0	48.0	80.0	1.020	1.0	0.0	NaN	normal	notpresent	notpresent	...	44	7800	5.2	yes	yes	no	good	no	no	ckd
1	1	7.0	50.0	1.020	4.0	0.0	NaN	normal	notpresent	notpresent	...	38	6000	NaN	no	no	no	good	no	no	ckd
2	2	62.0	80.0	1.010	2.0	3.0	normal	normal	notpresent	notpresent	...	31	7500	NaN	no	yes	no	poor	no	yes	ckd
3	3	48.0	70.0	1.005	4.0	0.0	normal	abnormal	present	notpresent	...	32	6700	3.9	yes	no	no	poor	yes	yes	ckd
4	4	51.0	80.0	1.010	2.0	0.0	normal	normal	notpresent	notpresent	...	35	7300	4.6	no	no	no	good	no	no	ckd

5 rows × 26 columns

In [5]:

```
df['wc']
```

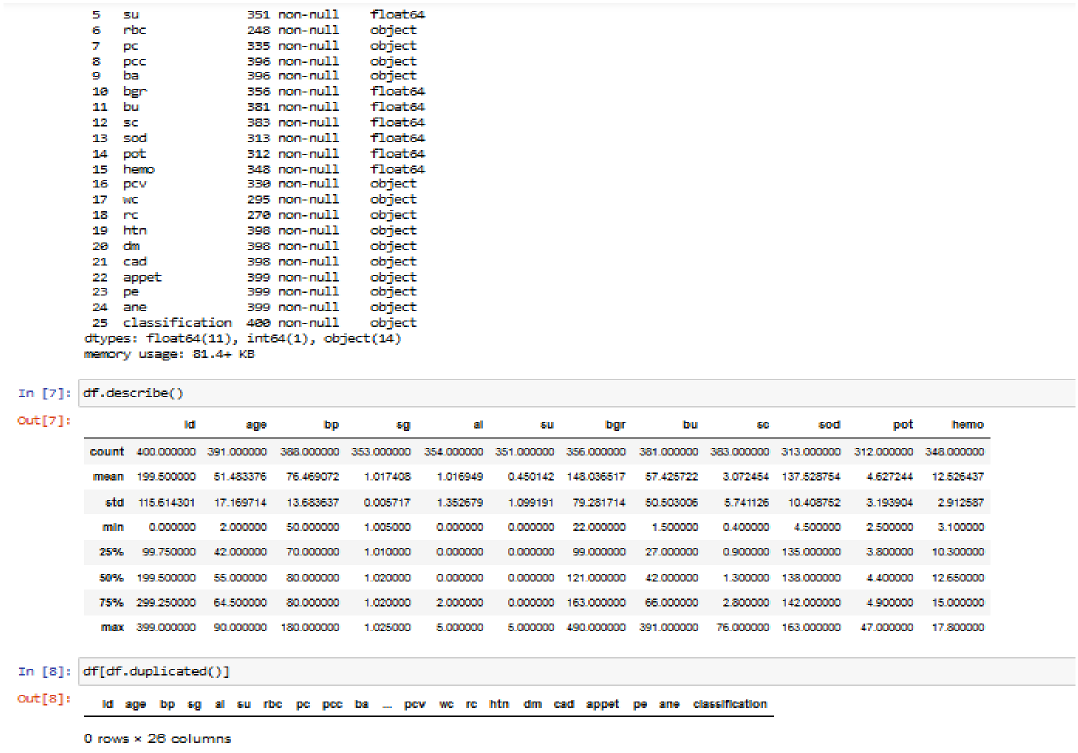
Out[5]:

```
0       7800
1       6000
2       7500
3       6700
4       7300
        ...
395     6700
396     7800
397     6600
398     7200
399     6800
Name: wc, Length: 400, dtype: object
```

In [6]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 400 entries, 0 to 399
Data columns (total 26 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   id              400 non-null    int64
 1   age             391 non-null    float64
 2   bp              388 non-null    float64
 3   sg              353 non-null    float64
 4   al              354 non-null    float64
```



```
 5   su              351 non-null    float64
 6   rbc             248 non-null    object
 7   pc              335 non-null    object
 8   pcc             396 non-null    object
 9   ba              396 non-null    object
 10  bgr             356 non-null    float64
 11  bu              381 non-null    float64
 12  sc              383 non-null    float64
 13  sod             313 non-null    float64
 14  pot             312 non-null    float64
 15  hemo            348 non-null    float64
 16  pcv             330 non-null    object
 17  wc              295 non-null    object
 18  rc              270 non-null    object
 19  htn             398 non-null    object
 20  dm              398 non-null    object
 21  cad             398 non-null    object
 22  appet           399 non-null    object
 23  pe              399 non-null    object
 24  ane             399 non-null    object
 25  classification  400 non-null    object
dtypes: float64(11), int64(1), object(14)
memory usage: 81.4+ KB
```

In [7]:
```
df.describe()
```

Out[7]:

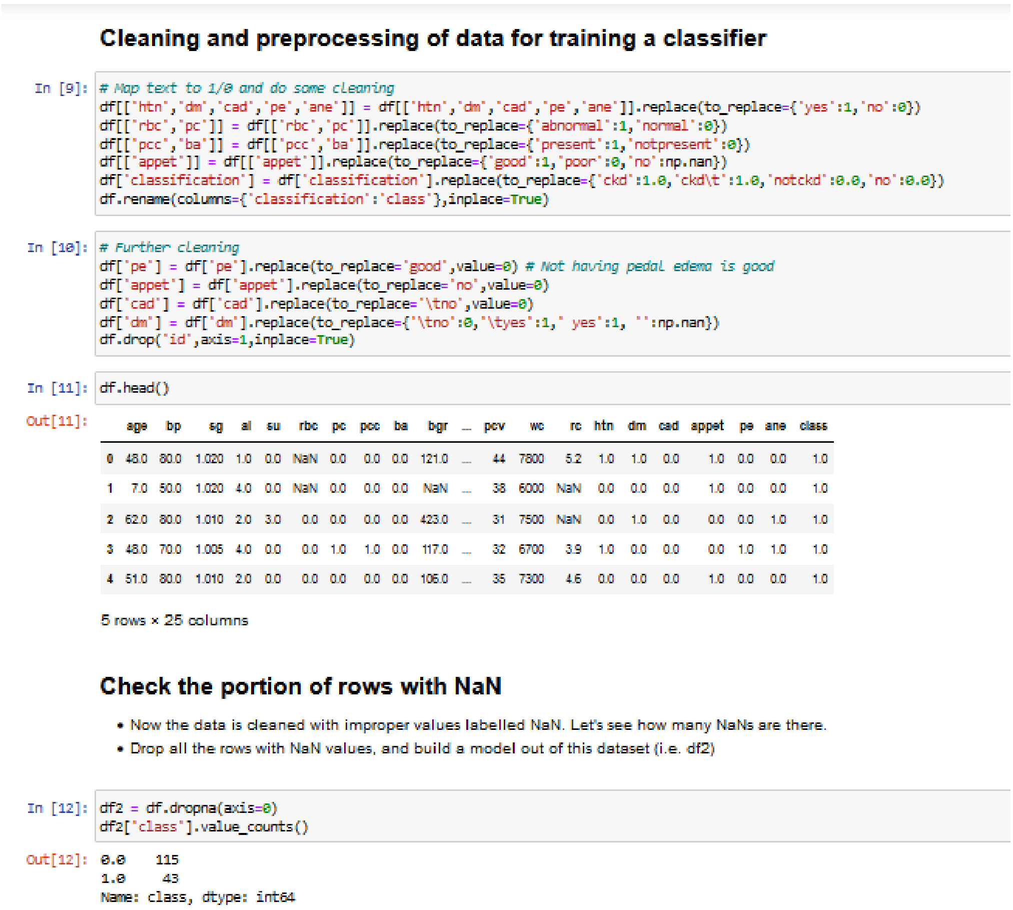
	id	age	bp	sg	al	su	bgr	bu	sc	sod	pot	hemo
count	400.000000	391.000000	388.000000	353.000000	354.000000	351.000000	356.000000	381.000000	383.000000	313.000000	312.000000	348.000000
mean	199.500000	51.483376	76.469072	1.017408	1.016949	0.450142	148.036517	57.425722	3.072454	137.528754	4.627244	12.526437
std	115.614301	17.169714	13.683637	0.005717	1.352679	1.099191	79.281714	50.503006	5.741126	10.408752	3.193904	2.912587
min	0.000000	2.000000	50.000000	1.005000	0.000000	0.000000	22.000000	1.500000	0.400000	4.500000	2.500000	3.100000
25%	99.750000	42.000000	70.000000	1.010000	0.000000	0.000000	99.000000	27.000000	0.900000	135.000000	3.800000	10.300000
50%	199.500000	55.000000	80.000000	1.020000	0.000000	0.000000	121.000000	42.000000	1.300000	138.000000	4.400000	12.650000
75%	299.250000	64.500000	80.000000	1.020000	2.000000	0.000000	163.000000	66.000000	2.800000	142.000000	4.900000	15.000000
max	399.000000	90.000000	180.000000	1.025000	5.000000	5.000000	490.000000	391.000000	76.000000	163.000000	47.000000	17.800000

In [8]:
```
df[df.duplicated()]
```

Out[8]:

id	age	bp	sg	al	su	rbc	pc	pcc	ba	...	pcv	wc	rc	htn	dm	cad	appet	pe	ane	classification

0 rows × 26 columns



Cleaning and preprocessing of data for training a classifier

In [9]:
```
# Map text to 1/0 and do some cleaning
df[['htn','dm','cad','pe','ane']] = df[['htn','dm','cad','pe','ane']].replace(to_replace={'yes':1,'no':0})
df[['rbc','pc']] = df[['rbc','pc']].replace(to_replace={'abnormal':1,'normal':0})
df[['pcc','ba']] = df[['pcc','ba']].replace(to_replace={'present':1,'notpresent':0})
df[['appet']] = df[['appet']].replace(to_replace={'good':1,'poor':0,'no':np.nan})
df['classification'] = df['classification'].replace(to_replace={'ckd':1.0,'ckd\t':1.0,'notckd':0.0,'no':0.0})
df.rename(columns={'classification':'class'},inplace=True)
```

In [10]:
```
# Further cleaning
df['pe'] = df['pe'].replace(to_replace='good',value=0) # Not having pedal edema is good
df['appet'] = df['appet'].replace(to_replace='no',value=0)
df['cad'] = df['cad'].replace(to_replace='\tno',value=0)
df['dm'] = df['dm'].replace(to_replace={'\tno':0,'\tyes':1,' yes':1, '':np.nan})
df.drop('id',axis=1,inplace=True)
```

In [11]:
```
df.head()
```

Out[11]:

	age	bp	sg	al	su	rbc	pc	pcc	ba	bgr	...	pcv	wc	rc	htn	dm	cad	appet	pe	ane	class
0	48.0	80.0	1.020	1.0	0.0	NaN	0.0	0.0	0.0	121.0	...	44	7800	5.2	1.0	1.0	0.0	1.0	0.0	0.0	1.0
1	7.0	50.0	1.020	4.0	0.0	NaN	0.0	0.0	0.0	NaN	...	38	6000	NaN	0.0	0.0	0.0	1.0	0.0	0.0	1.0
2	62.0	80.0	1.010	2.0	3.0	0.0	0.0	0.0	0.0	423.0	...	31	7500	NaN	0.0	1.0	0.0	0.0	0.0	1.0	1.0
3	48.0	70.0	1.005	4.0	0.0	0.0	1.0	1.0	0.0	117.0	...	32	6700	3.9	1.0	0.0	0.0	0.0	1.0	1.0	1.0
4	51.0	80.0	1.010	2.0	0.0	0.0	0.0	0.0	0.0	106.0	...	35	7300	4.6	0.0	0.0	0.0	1.0	0.0	0.0	1.0

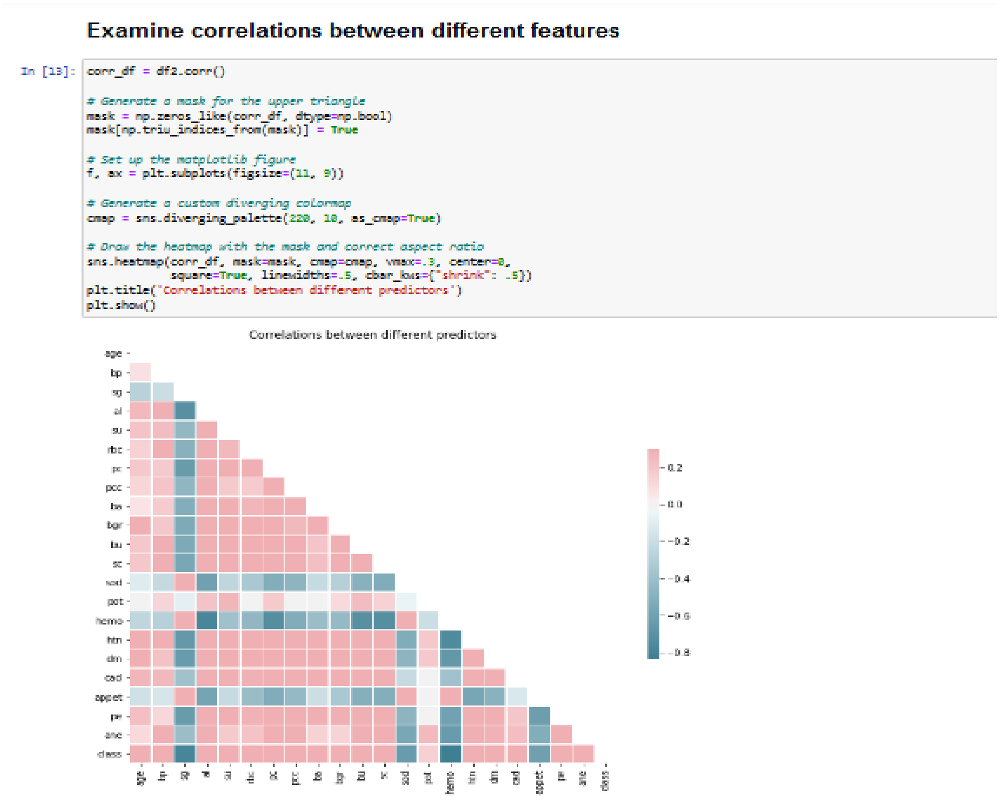
5 rows × 25 columns

Check the portion of rows with NaN

- Now the data is cleaned with improper values labelled NaN. Let's see how many NaNs are there.
- Drop all the rows with NaN values, and build a model out of this dataset (i.e. df2)

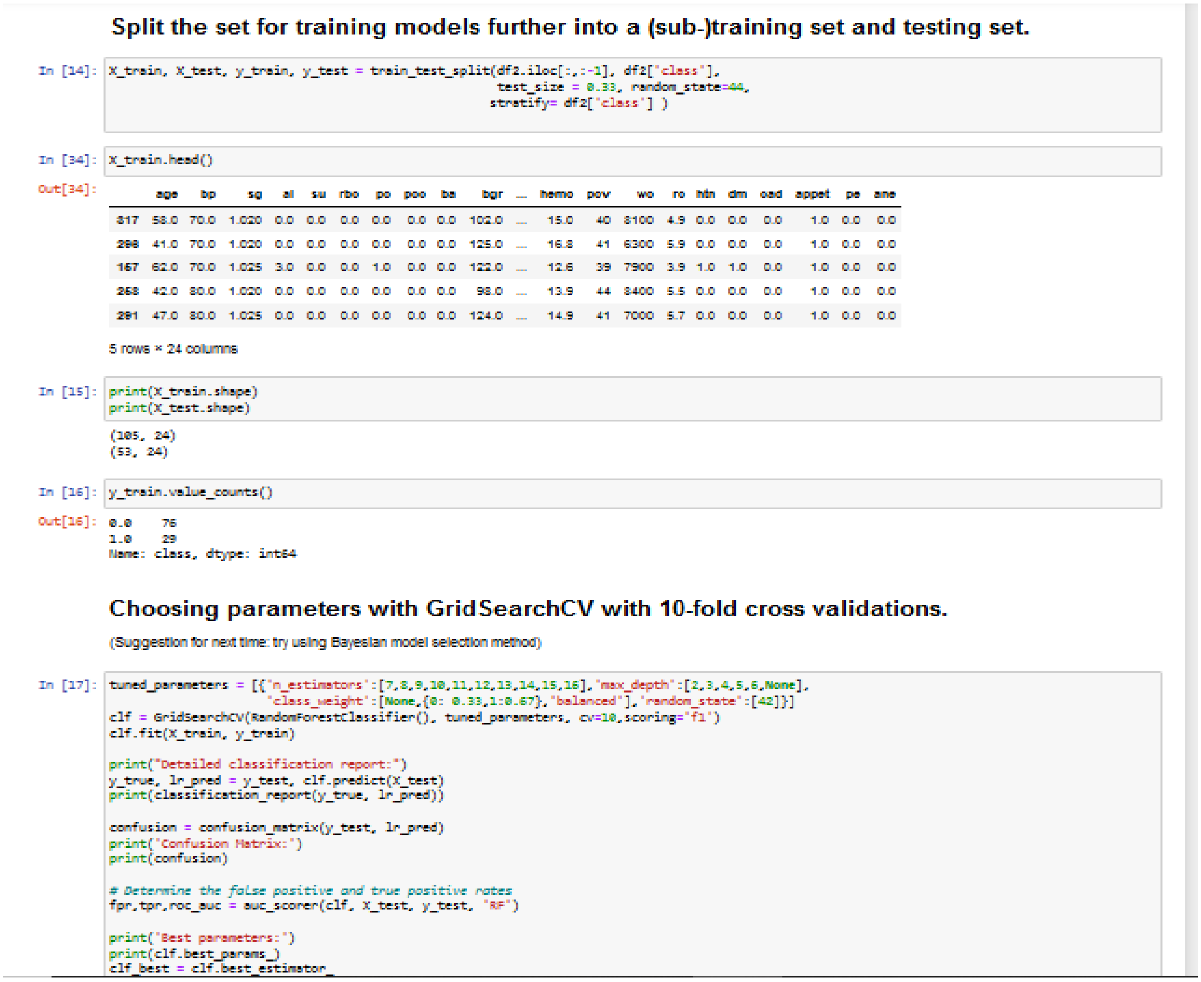
In [12]:
```
df2 = df.dropna(axis=0)
df2['class'].value_counts()
```

Out[12]:
```
0.0    115
1.0     43
Name: class, dtype: int64
```



Examine correlations between different features

```
In [13]: corr_df = df2.corr()

         # Generate a mask for the upper triangle
         mask = np.zeros_like(corr_df, dtype=np.bool)
         mask[np.triu_indices_from(mask)] = True

         # Set up the matplotlib figure
         f, ax = plt.subplots(figsize=(11, 9))

         # Generate a custom diverging colormap
         cmap = sns.diverging_palette(220, 10, as_cmap=True)

         # Draw the heatmap with the mask and correct aspect ratio
         sns.heatmap(corr_df, mask=mask, cmap=cmap, vmax=.3, center=0,
                     square=True, linewidths=.5, cbar_kws={"shrink": .5})
         plt.title('Correlations between different predictors')
         plt.show()
```



Split the set for training models further into a (sub-)training set and testing set.

```
In [14]: X_train, X_test, y_train, y_test = train_test_split(df2.iloc[:,:-1], df2['class'],
                                                              test_size = 0.33, random_state=44,
                                                             stratify= df2['class'] )
```

```
In [34]: X_train.head()
```

Out[34]:

	age	bp	sg	al	su	rbc	pc	pcc	ba	bgr	...	hemo	pcv	wc	rc	htn	dm	cad	appet	pe	ane
317	58.0	70.0	1.020	0.0	0.0	0.0	0.0	0.0	0.0	102.0	...	15.0	40	8100	4.9	0.0	0.0	0.0	1.0	0.0	0.0
298	41.0	70.0	1.020	0.0	0.0	0.0	0.0	0.0	0.0	125.0	...	16.8	41	6300	5.9	0.0	0.0	0.0	1.0	0.0	0.0
167	62.0	70.0	1.025	3.0	0.0	0.0	1.0	0.0	0.0	122.0	...	12.6	39	7900	3.9	1.0	1.0	0.0	1.0	0.0	0.0
268	42.0	80.0	1.020	0.0	0.0	0.0	0.0	0.0	0.0	98.0	...	13.9	44	8400	5.5	0.0	0.0	0.0	1.0	0.0	0.0
281	47.0	80.0	1.025	0.0	0.0	0.0	0.0	0.0	0.0	124.0	...	14.9	41	7000	5.7	0.0	0.0	0.0	1.0	0.0	0.0

5 rows × 24 columns

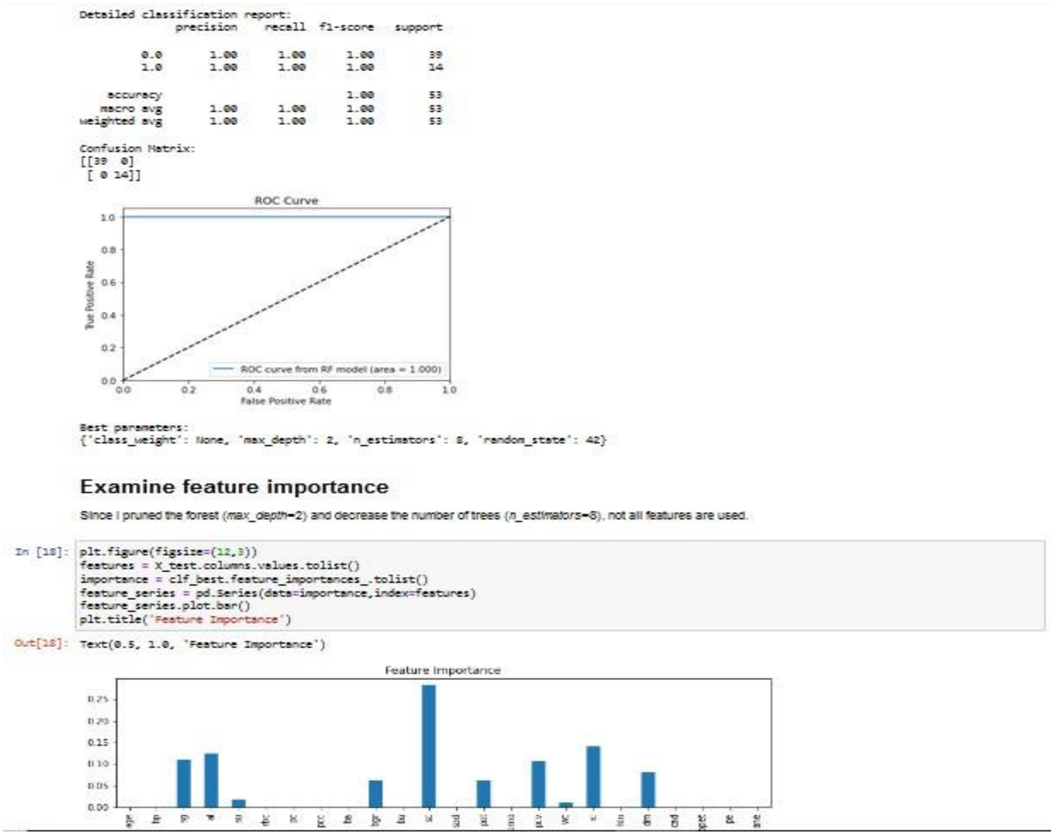
```
In [15]: print(X_train.shape)
         print(X_test.shape)

         (105, 24)
         (53, 24)
```

```
In [16]: y_train.value_counts()

Out[16]: 0.0    76
         1.0    29
         Name: class, dtype: int64
```

Choosing parameters with GridSearchCV with 10-fold cross validations.

(Suggestion for next time: try using Bayesian model selection method)

```
In [17]: tuned_parameters = [{'n_estimators':[7,8,9,10,11,12,13,14,15,16],'max_depth':[2,3,4,5,6,None],
                             'class_weight':[None,{0: 0.33,1:0.67},'balanced'],'random_state':[42]}]
         clf = GridSearchCV(RandomForestClassifier(), tuned_parameters, cv=10,scoring='f1')
         clf.fit(X_train, y_train)

         print("Detailed classification report:")
         y_true, lr_pred = y_test, clf.predict(X_test)
         print(classification_report(y_true, lr_pred))

         confusion = confusion_matrix(y_test, lr_pred)
         print('Confusion Matrix:')
         print(confusion)

         # Determine the false positive and true positive rates
         fpr,tpr,roc_auc = auc_scorer(clf, X_test, y_test, 'RF')

         print('Best parameters:')
         print(clf.best_params_)
         clf_best = clf.best_estimator_
```



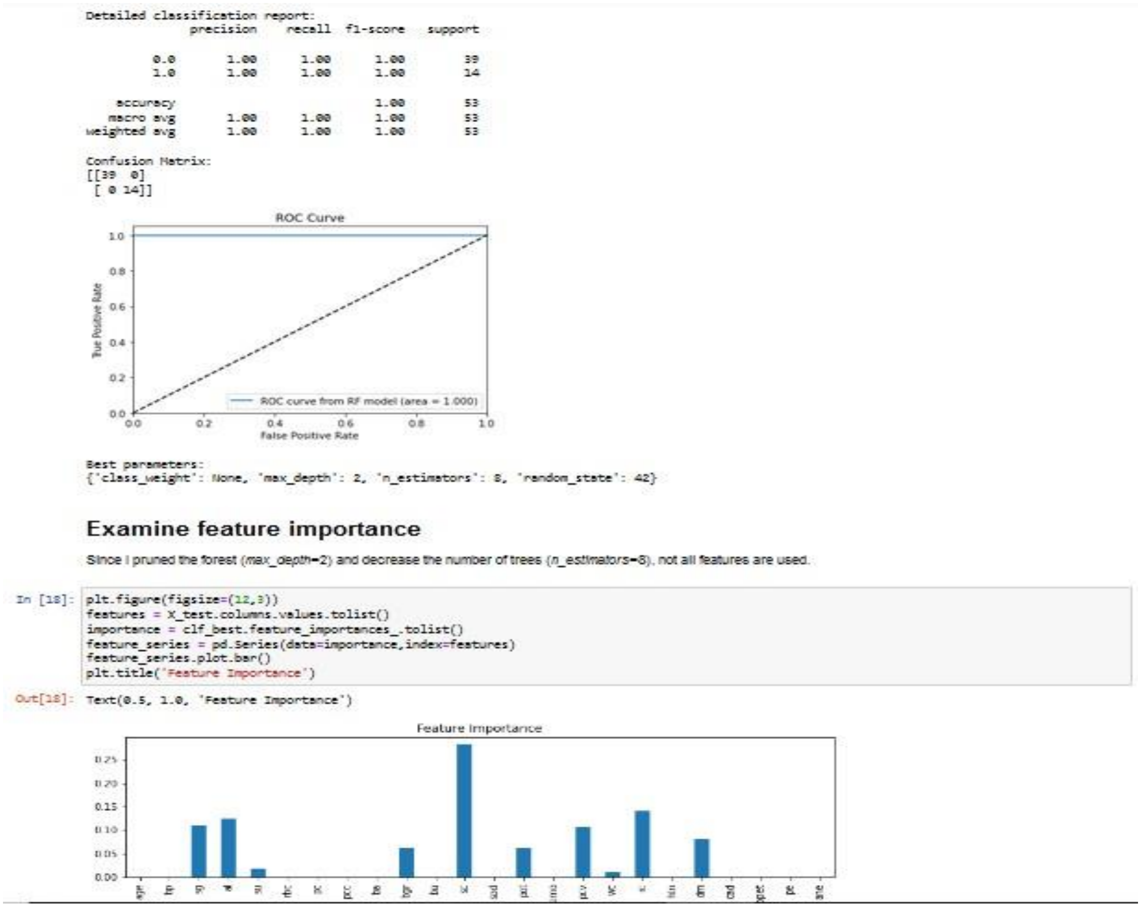
```
Detailed classification report:
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00        39
         1.0       1.00      1.00      1.00        14

    accuracy                           1.00        53
   macro avg       1.00      1.00      1.00        53
weighted avg       1.00      1.00      1.00        53

Confusion Matrix:
[[39  0]
 [ 0 14]]
```

```
Best parameters:
{'class_weight': None, 'max_depth': 2, 'n_estimators': 8, 'random_state': 42}
```

Examine feature importance

Since I pruned the forest (*max_depth*=2) and decrease the number of trees (*n_estimators*=8), not all features are used.

```
In [18]: plt.figure(figsize=(12,3))
         features = X_test.columns.values.tolist()
         importance = clf_best.feature_importances_.tolist()
         feature_series = pd.Series(data=importance,index=features)
         feature_series.plot.bar()
         plt.title('Feature Importance')

Out[18]: Text(0.5, 1.0, 'Feature Importance')
```



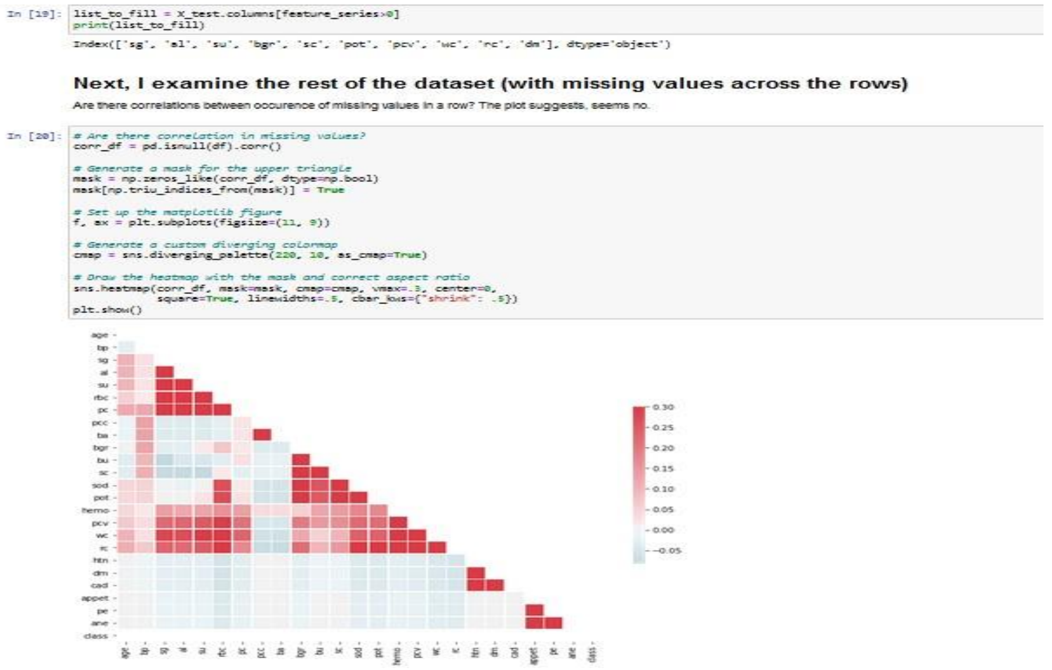
```
Detailed classification report:
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00        39
         1.0       1.00      1.00      1.00        14

    accuracy                           1.00        53
   macro avg       1.00      1.00      1.00        53
weighted avg       1.00      1.00      1.00        53

Confusion Matrix:
[[39  0]
 [ 0 14]]
```

```
Best parameters:
{'class_weight': None, 'max_depth': 2, 'n_estimators': 8, 'random_state': 42}
```

Examine feature importance

Since I pruned the forest (max_depth=2) and decrease the number of trees (n_estimators=8), not all features are used.

```
In [18]: plt.figure(figsize=(12,3))
         features = X_test.columns.values.tolist()
         importance = clf_best.feature_importances_.tolist()
         feature_series = pd.Series(data=importance,index=features)
         feature_series.plot.bar()
         plt.title('Feature Importance')

Out[18]: Text(0.5, 1.0, 'Feature Importance')
```

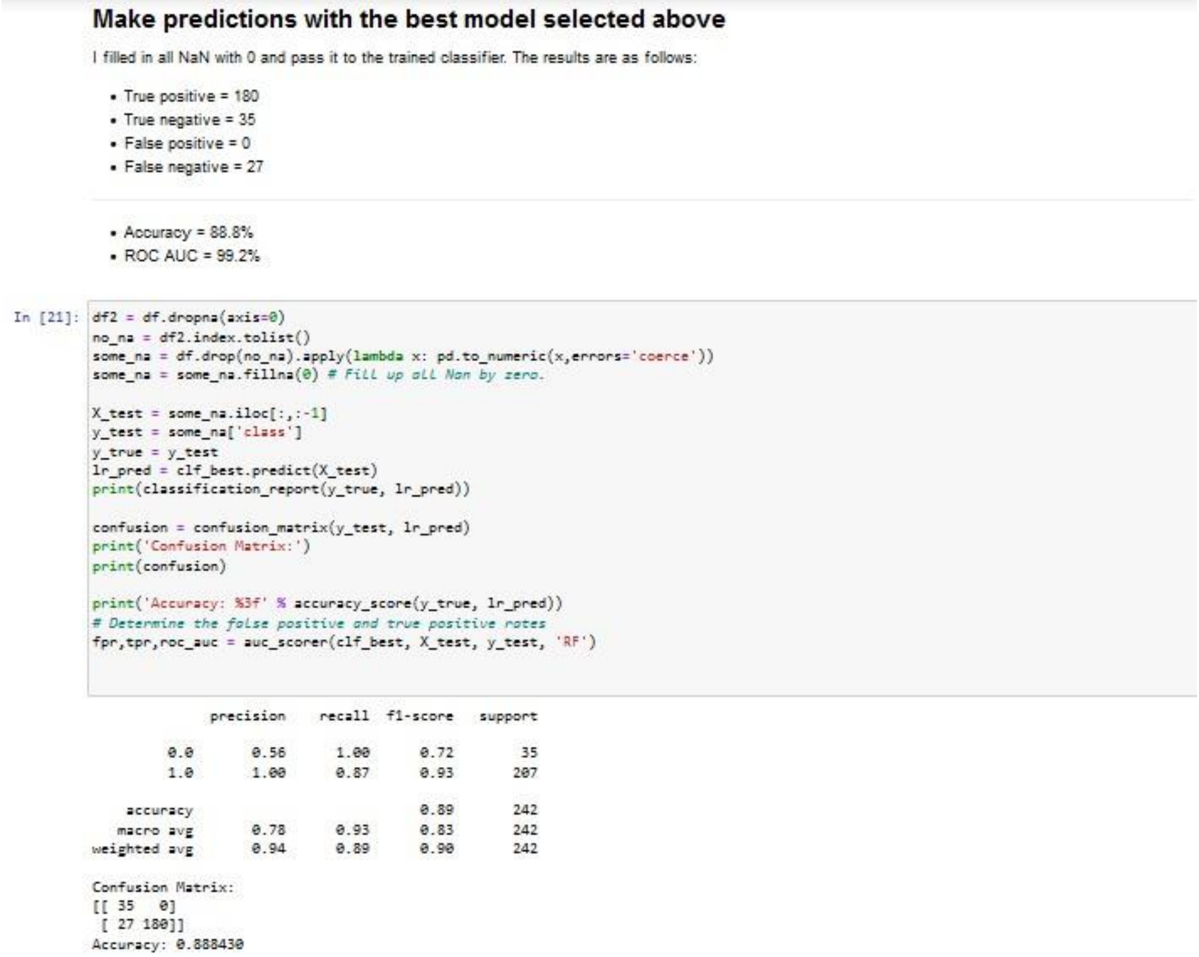


```
In [19]: list_to_fill = X_test.columns[feature_series>0]
         print(list_to_fill)

         Index(['sg', 'al', 'su', 'bgr', 'sc', 'pot', 'pcv', 'wc', 'rc', 'dm'], dtype='object')
```

Next, I examine the rest of the dataset (with missing values across the rows)

Are there correlations between occurence of missing values in a row? The plot suggests, seems no.

```
In [20]: # Are there correlation in missing values?
         corr_df = pd.isnull(df).corr()

         # Generate a mask for the upper triangle
         mask = np.zeros_like(corr_df, dtype=np.bool)
         mask[np.triu_indices_from(mask)] = True

         # Set up the matplotlib figure
         f, ax = plt.subplots(figsize=(11, 9))

         # Generate a custom diverging colormap
         cmap = sns.diverging_palette(220, 10, as_cmap=True)

         # Draw the heatmap with the mask and correct aspect ratio
         sns.heatmap(corr_df, mask=mask, cmap=cmap, vmax=.3, center=0,
                     square=True, linewidths=.5, cbar_kws={"shrink": .5})
         plt.show()
```



Make predictions with the best model selected above

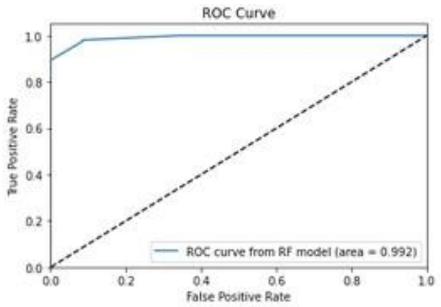
I filled in all NaN with 0 and pass it to the trained classifier. The results are as follows:

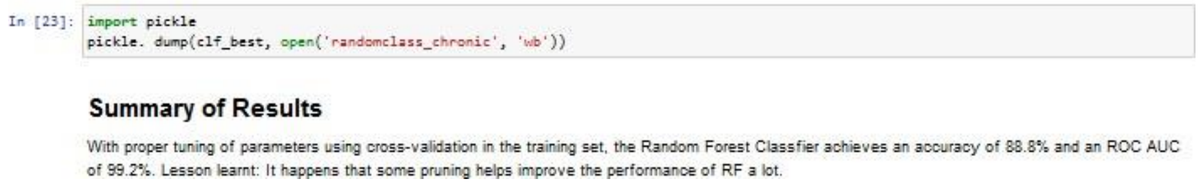
- True positive = 180
- True negative = 35
- False positive = 0
- False negative = 27

- Accuracy = 88.8%
- ROC AUC = 99.2%

```
In [21]: df2 = df.dropna(axis=0)
         no_na = df2.index.tolist()
         some_na = df.drop(no_na).apply(lambda x: pd.to_numeric(x,errors='coerce'))
         some_na = some_na.fillna(0) # Fill up all Nan by zero.

         X_test = some_na.iloc[:,:-1]
         y_test = some_na['class']
         y_true = y_test
         lr_pred = clf_best.predict(X_test)
         print(classification_report(y_true, lr_pred))

         confusion = confusion_matrix(y_test, lr_pred)
         print('Confusion Matrix:')
         print(confusion)

         print('Accuracy: %3f' % accuracy_score(y_true, lr_pred))
         # Determine the false positive and true positive rates
         fpr,tpr,roc_auc = auc_scorer(clf_best, X_test, y_test, 'RF')
```

```
              precision    recall  f1-score   support

         0.0       0.56      1.00      0.72        35
         1.0       1.00      0.87      0.93       207

    accuracy                           0.89       242
   macro avg       0.78      0.93      0.83       242
weighted avg       0.94      0.89      0.90       242

Confusion Matrix:
[[ 35   0]
 [ 27 180]]
Accuracy: 0.888430
```





```
In [23]: import pickle
         pickle. dump(clf_best, open('randomclass_chronic', 'wb'))
```

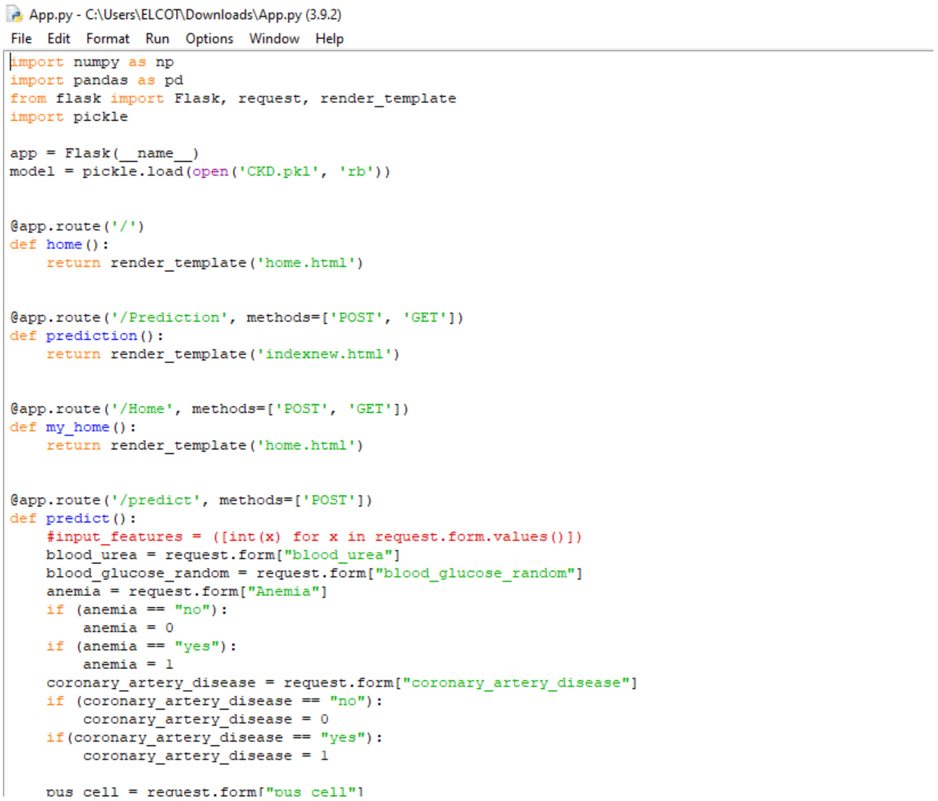
Summary of Results

With proper tuning of parameters using cross-validation in the training set, the Random Forest Classfier achieves an accuracy of 88.8% and an ROC AUC of 99.2%. Lesson learnt: It happens that some pruning helps improve the performance of RF a lot.

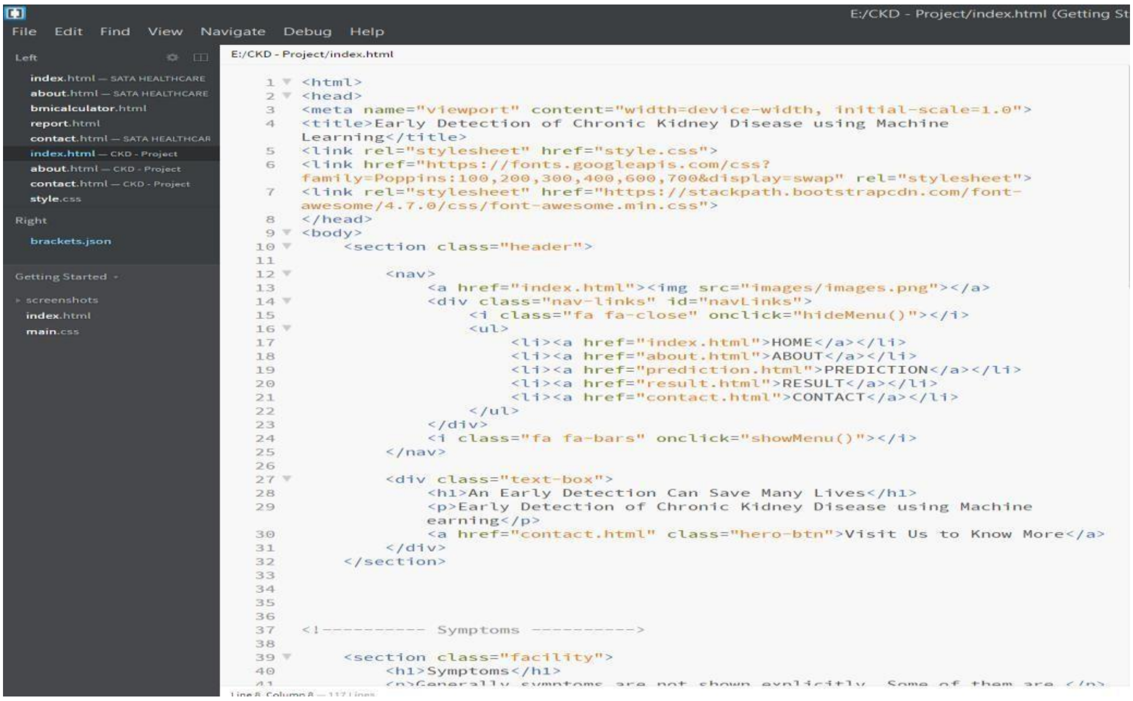
7.2 FEATURE 2:

During Sprint2 we have planned for Creating HTML files, Build Python code and run the app1.
Building flask file:
app.py screen shots

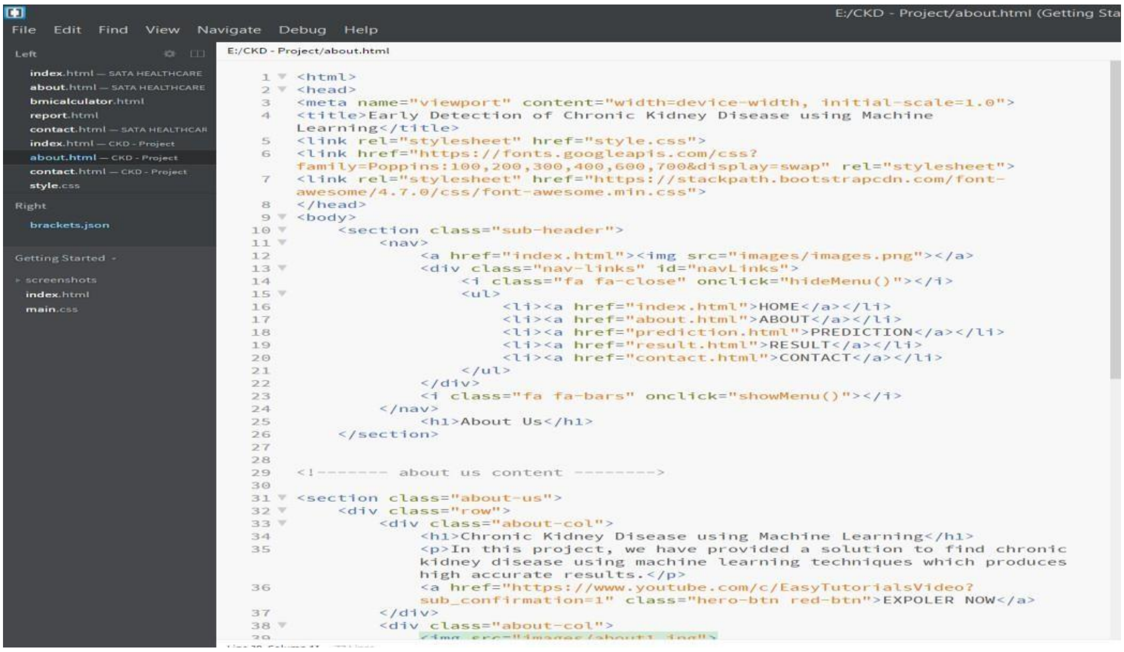
- **App.py Code Screen**


App.py - C:\Users\ELCOT\Downloads\App.py (3.9.2)
File Edit Format Run Options Window Help

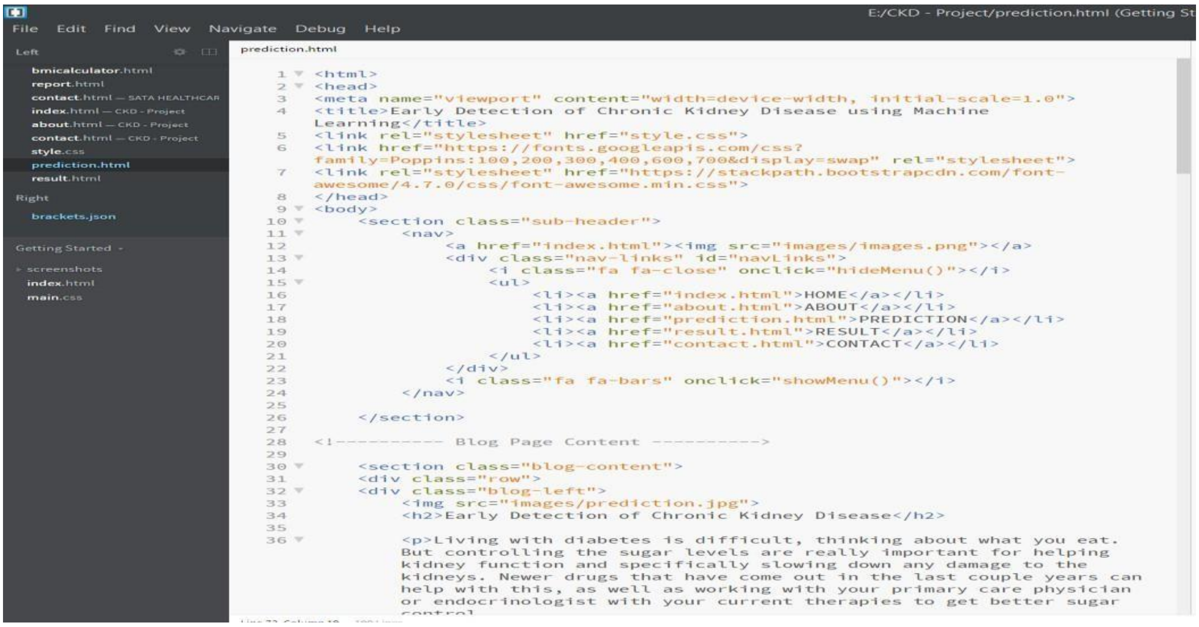
```
import numpy as np
import pandas as pd
from flask import Flask, request, render_template
import pickle

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


@app.route('/')
def home():
    return render_template('home.html')


@app.route('/Prediction', methods=['POST', 'GET'])
def prediction():
    return render_template('indexnew.html')


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


@app.route('/predict', methods=['POST'])
def predict():
    #input_features = ([int(x) for x in request.form.values()])
    blood_urea = request.form["blood_urea"]
    blood_glucose_random = request.form["blood_glucose_random"]
    anemia = request.form["Anemia"]
    if (anemia == "no"):
        anemia = 0
    if (anemia == "yes"):
        anemia = 1
    coronary_artery_disease = request.form["coronary_artery_disease"]
    if (coronary_artery_disease == "no"):
        coronary_artery_disease = 0
    if(coronary_artery_disease == "yes"):
        coronary_artery_disease = 1

    pus_cell = request.form["pus_cell"]
```


E:/CKD - Project/index.html (Getting St
File
Edit
Find
View
Navigate
Debug
Help
Left
index.html — SATA HEALTHCARE
about.html — SATA HEALTHCARE
bmicalculator.html
report.html
contact.html — SATA HEALTHCAR
index.html — CKD - Project
about.html — CKD - Project
contact.html — CKD - Project
style.css
Right
brackets.json
Getting Started
screenshots
index.html
main.css
E:/CKD - Project/index.html
Line 8, Column 8 — 117 Lines

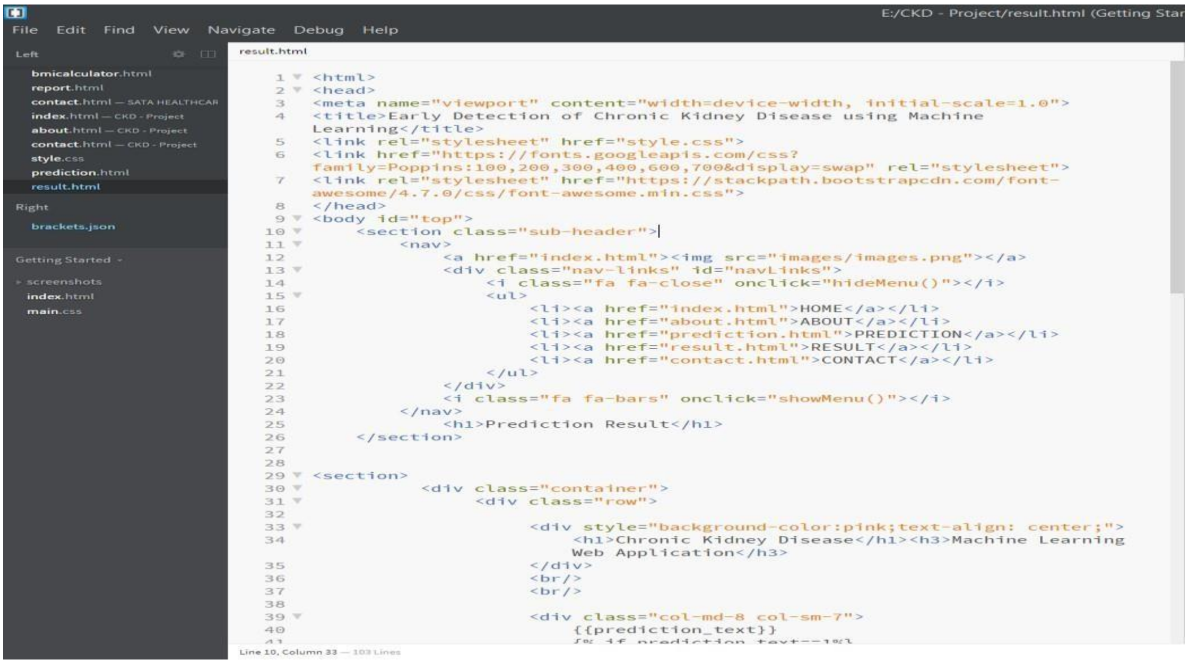
```
<html>
<head>
<meta name="viewport" content="width=device-width, initial-scale=1.0">
<title>Early Detection of Chronic Kidney Disease using Machine
Learning</title>
<link rel="stylesheet" href="style.css">
<link href="https://fonts.googleapis.com/css?
family=Poppins:100,200,300,400,600,700&display=swap" rel="stylesheet">
<link rel="stylesheet" href="https://stackpath.bootstrapcdn.com/font-
awesome/4.7.0/css/font-awesome.min.css">
</head>
<body>
    <section class="header">

        <nav>
            <a href="index.html"><img src="images/images.png"></a>
            <div class="nav-links" id="navLinks">
                <i class="fa fa-close" onclick="hideMenu()"></i>
                <ul>
                    <li><a href="index.html">HOME</a></li>
                    <li><a href="about.html">ABOUT</a></li>
                    <li><a href="prediction.html">PREDICTION</a></li>
                    <li><a href="result.html">RESULT</a></li>
                    <li><a href="contact.html">CONTACT</a></li>
                </ul>
            </div>
            <i class="fa fa-bars" onclick="showMenu()"></i>
        </nav>

        <div class="text-box">
            <h1>An Early Detection Can Save Many Lives</h1>
            <p>Early Detection of Chronic Kidney Disease using Machine
            earning</p>
            <a href="contact.html" class="hero-btn">Visit Us to Know More</a>
        </div>
    </section>




<!---------- Symptoms ---------->

    <section class="facility">
        <h1>Symptoms</h1>
        <p>Generally symptoms are not shown explicitly. Some of them are </p>
```


E:/CKD - Project/about.html (Getting Sta
File
Edit
Find
View
Navigate
Debug
Help
Left
index.html — SATA HEALTHCARE
about.html — SATA HEALTHCARE
bmicalculator.html
report.html
contact.html — SATA HEALTHCAR
index.html — CKD - Project
about.html — CKD - Project
contact.html — CKD - Project
style.css
Right
brackets.json
Getting Started
screenshots
index.html
main.css
E:/CKD - Project/about.html

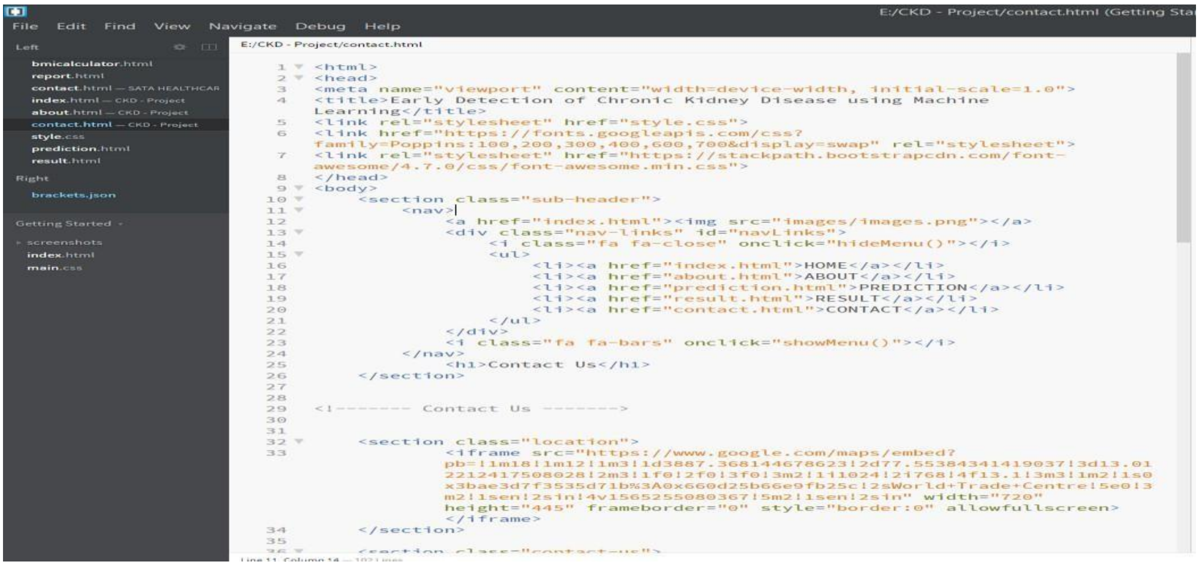
```
<html>
<head>
<meta name="viewport" content="width=device-width, initial-scale=1.0">
<title>Early Detection of Chronic Kidney Disease using Machine
Learning</title>
<link rel="stylesheet" href="style.css">
<link href="https://fonts.googleapis.com/css?
family=Poppins:100,200,300,400,600,700&display=swap" rel="stylesheet">
<link rel="stylesheet" href="https://stackpath.bootstrapcdn.com/font-
awesome/4.7.0/css/font-awesome.min.css">
</head>
<body>
    <section class="sub-header">
        <nav>
            <a href="index.html"><img src="images/images.png"></a>
            <div class="nav-links" id="navLinks">
                <i class="fa fa-close" onclick="hideMenu()"></i>
                <ul>
                    <li><a href="index.html">HOME</a></li>
                    <li><a href="about.html">ABOUT</a></li>
                    <li><a href="prediction.html">PREDICTION</a></li>
                    <li><a href="result.html">RESULT</a></li>
                    <li><a href="contact.html">CONTACT</a></li>
                </ul>
            </div>
            <i class="fa fa-bars" onclick="showMenu()"></i>
        </nav>
            <h1>About Us</h1>
    </section>


<!------- about us content --------->

<section class="about-us">
    <div class="row">
        <div class="about-col">
            <h1>Chronic Kidney Disease using Machine Learning</h1>
            <p>In this project, we have provided a solution to find chronic
            kidney disease using machine learning techniques which produces
            high accurate results.</p>
            <a href="https://www.youtube.com/c/EasyTutorialsVideo?
            sub_confirmation=1" class="hero-btn red-btn">EXPOLER NOW</a>
        </div>
        <div class="about-col">
            <img src="images/about1.jpg">
```



```html
<html>
<head>
<meta name="viewport" content="width=device-width, initial-scale=1.0">
<title>Early Detection of Chronic Kidney Disease using Machine
Learning</title>
<link rel="stylesheet" href="style.css">
<link href="https://fonts.googleapis.com/css?
family=Poppins:100,200,300,400,600,700&display=swap" rel="stylesheet">
<link rel="stylesheet" href="https://stackpath.bootstrapcdn.com/font-
awesome/4.7.0/css/font-awesome.min.css">
</head>
<body>
    <section class="sub-header">
        <nav>
            <a href="index.html"><img src="images/images.png"></a>
            <div class="nav-links" id="navLinks">
                <i class="fa fa-close" onclick="hideMenu()"></i>
                <ul>
                    <li><a href="index.html">HOME</a></li>
                    <li><a href="about.html">ABOUT</a></li>
                    <li><a href="prediction.html">PREDICTION</a></li>
                    <li><a href="result.html">RESULT</a></li>
                    <li><a href="contact.html">CONTACT</a></li>
                </ul>
            </div>
            <i class="fa fa-bars" onclick="showMenu()"></i>
        </nav>

    </section>

<!---------- Blog Page Content ---------->

    <section class="blog-content">
    <div class="row">
    <div class="blog-left">
        <img src="images/prediction.jpg">
        <h2>Early Detection of Chronic Kidney Disease</h2>

        <p>Living with diabetes is difficult, thinking about what you eat.
        But controlling the sugar levels are really important for helping
        kidney function and specifically slowing down any damage to the
        kidneys. Newer drugs that have come out in the last couple years can
        help with this, as well as working with your primary care physician
        or endocrinologist with your current therapies to get better sugar
        control
```


E:/CKD - Project/result.html (Getting Star
File
Edit
Find
View
Navigate
Debug
Help
Left
bmicalculator.html
report.html
contact.html — SATA HEALTHCAR
index.html — CKD - Project
about.html — CKD - Project
contact.html — CKD - Project
style.css
prediction.html
result.html
Right
brackets.json
Getting Started
screenshots
index.html
main.css
result.html
Line 10, Column 33 — 103 Lines

```html
<html>
<head>
<meta name="viewport" content="width=device-width, initial-scale=1.0">
<title>Early Detection of Chronic Kidney Disease using Machine
Learning</title>
<link rel="stylesheet" href="style.css">
<link href="https://fonts.googleapis.com/css?
family=Poppins:100,200,300,400,600,700&display=swap" rel="stylesheet">
<link rel="stylesheet" href="https://stackpath.bootstrapcdn.com/font-
awesome/4.7.0/css/font-awesome.min.css">
</head>
<body id="top">
    <section class="sub-header">
        <nav>
            <a href="index.html"><img src="images/images.png"></a>
            <div class="nav-links" id="navLinks">
                <i class="fa fa-close" onclick="hideMenu()"></i>
                <ul>
                    <li><a href="index.html">HOME</a></li>
                    <li><a href="about.html">ABOUT</a></li>
                    <li><a href="prediction.html">PREDICTION</a></li>
                    <li><a href="result.html">RESULT</a></li>
                    <li><a href="contact.html">CONTACT</a></li>
                </ul>
            </div>
            <i class="fa fa-bars" onclick="showMenu()"></i>
        </nav>
            <h1>Prediction Result</h1>
    </section>


<section>
          <div class="container">
                <div class="row">

                    <div style="background-color:pink;text-align: center;">
                        <h1>Chronic Kidney Disease</h1><h3>Machine Learning
                        Web Application</h3>
                    </div>
                    <br/>
                    <br/>

                    <div class="col-md-8 col-sm-7">
                        {{prediction_text}}
                        {% if prediction_text==1%}
```

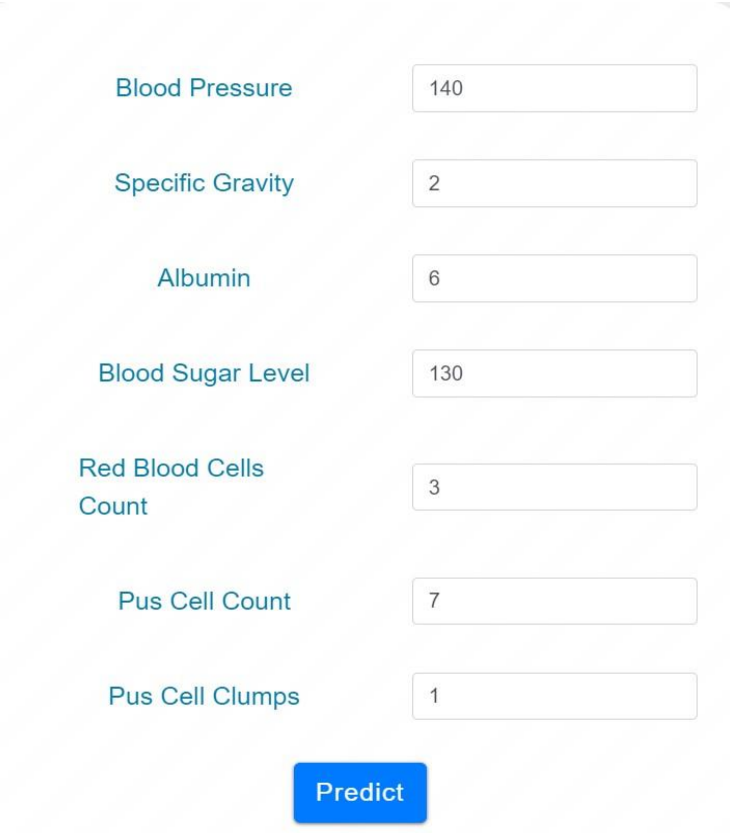

E:/CKD - Project/contact.html (Getting Sta
File
Edit
Find
View
Navigate
Debug
Help
Left
bmicalculator.html
report.html
contact.html — SATA HEALTHCAR
index.html — CKD - Project
about.html — CKD - Project
contact.html — CKD - Project
style.css
prediction.html
result.html
Right
brackets.json
Getting Started
screenshots
index.html
main.css
E:/CKD - Project/contact.html

```
<html>
<head>
<meta name="viewport" content="width=device-width, initial-scale=1.0">
<title>Early Detection of Chronic Kidney Disease using Machine
Learning</title>
<link rel="stylesheet" href="style.css">
<link href="https://fonts.googleapis.com/css?
family=Poppins:100,200,300,400,600,700&display=swap" rel="stylesheet">
<link rel="stylesheet" href="https://stackpath.bootstrapcdn.com/font-
awesome/4.7.0/css/font-awesome.min.css">
</head>
<body>
    <section class="sub-header">
        <nav>
            <a href="index.html"><img src="images/images.png"></a>
            <div class="nav-links" id="navLinks">
                <i class="fa fa-close" onclick="hideMenu()"></i>
                <ul>
                    <li><a href="index.html">HOME</a></li>
                    <li><a href="about.html">ABOUT</a></li>
                    <li><a href="prediction.html">PREDICTION</a></li>
                    <li><a href="result.html">RESULT</a></li>
                    <li><a href="contact.html">CONTACT</a></li>
                </ul>
            </div>
            <i class="fa fa-bars" onclick="showMenu()"></i>
        </nav>
            <h1>Contact Us</h1>
    </section>


<!------- Contact Us -------->


    <section class="location">
            <iframe src="https://www.google.com/maps/embed?
            pb=!1m18!1m12!1m3!1d3887.368144678623!2d77.55384341419037!3d13.01
            2212417508028!2m3!1f0!2f0!3f0!3m2!1i1024!2i768!4f13.1!3m3!1m2!1s0
            x3bae3d7f3535d71b%3A0x660d25b66e9fb25c!2sWorld+Trade+Centre!5e0!3
            m2!1sen!2sin!4v1565255080367!5m2!1sen!2sin" width="720"
            height="445" frameborder="0" style="border:0" allowfullscreen>
            </iframe>
    </section>

    <section class="contact-us">
```

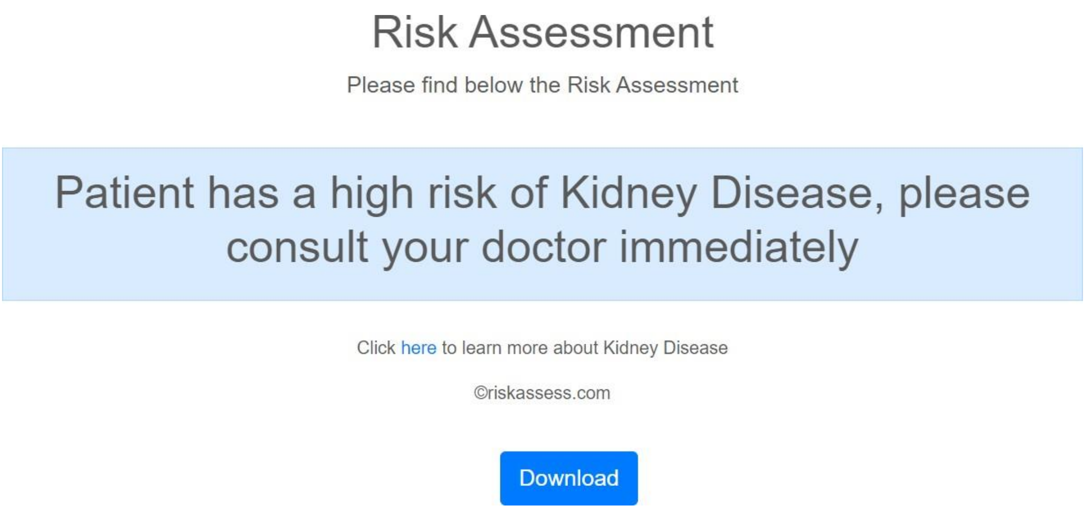
8. TESTING

8.1 Test Cases:

CKD:



Blood Pressure	140
Specific Gravity	2
Albumin	6
Blood Sugar Level	130
Red Blood Cells Count	3
Pus Cell Count	7
Pus Cell Clumps	1



Risk Assessment

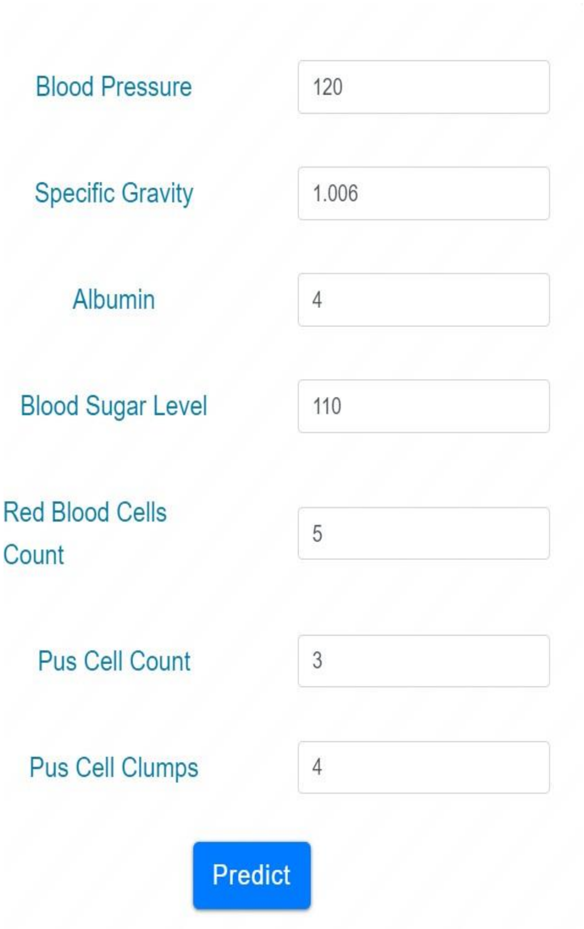
Please find below the Risk Assessment

Patient has a high risk of Kidney Disease, please consult your doctor immediately

Click here to learn more about Kidney Disease

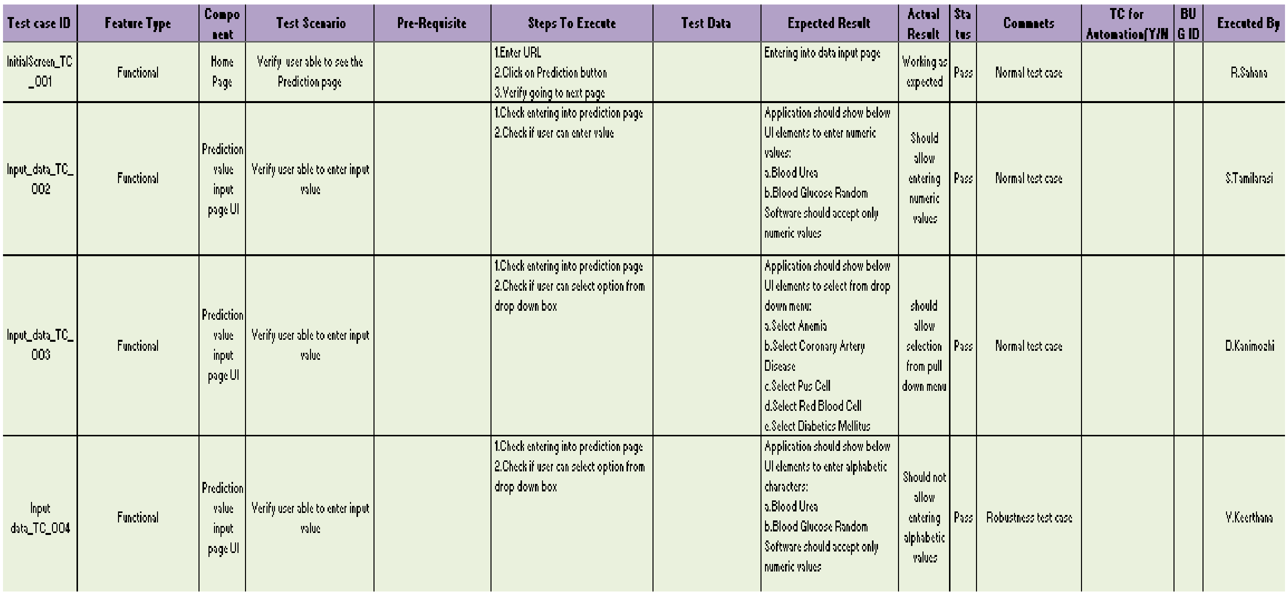
©riskassess.com

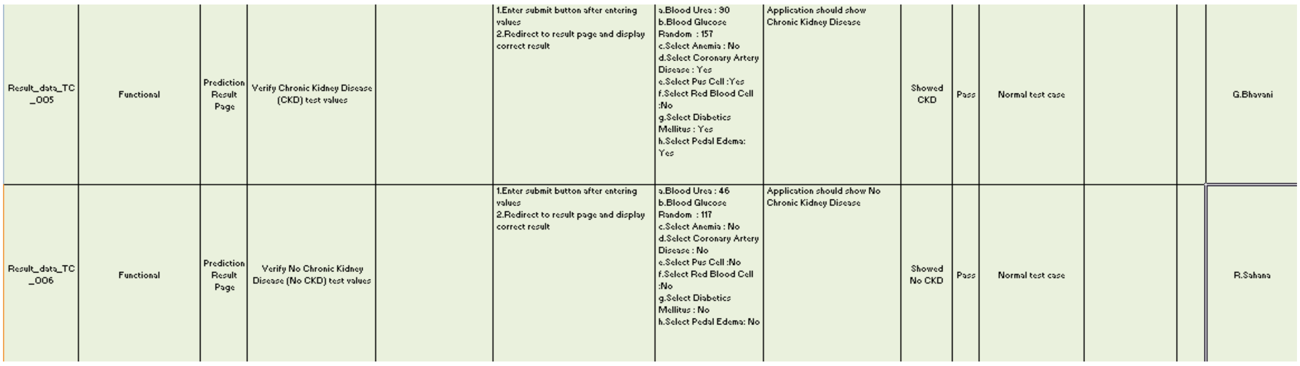
No CKD:

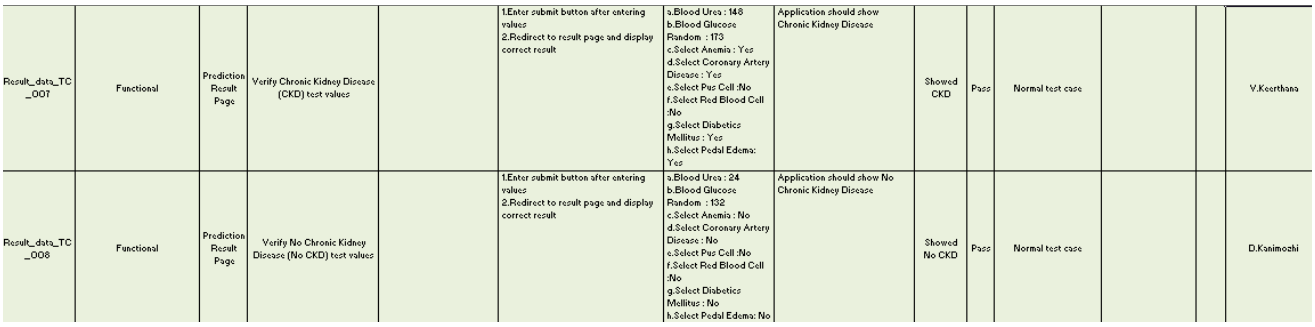


Blood Pressure	120
Specific Gravity	1.006
Albumin	4
Blood Sugar Level	110
Red Blood Cells Count	5
Pus Cell Count	3
Pus Cell Clumps	4

8.2 User Acceptance Testing:



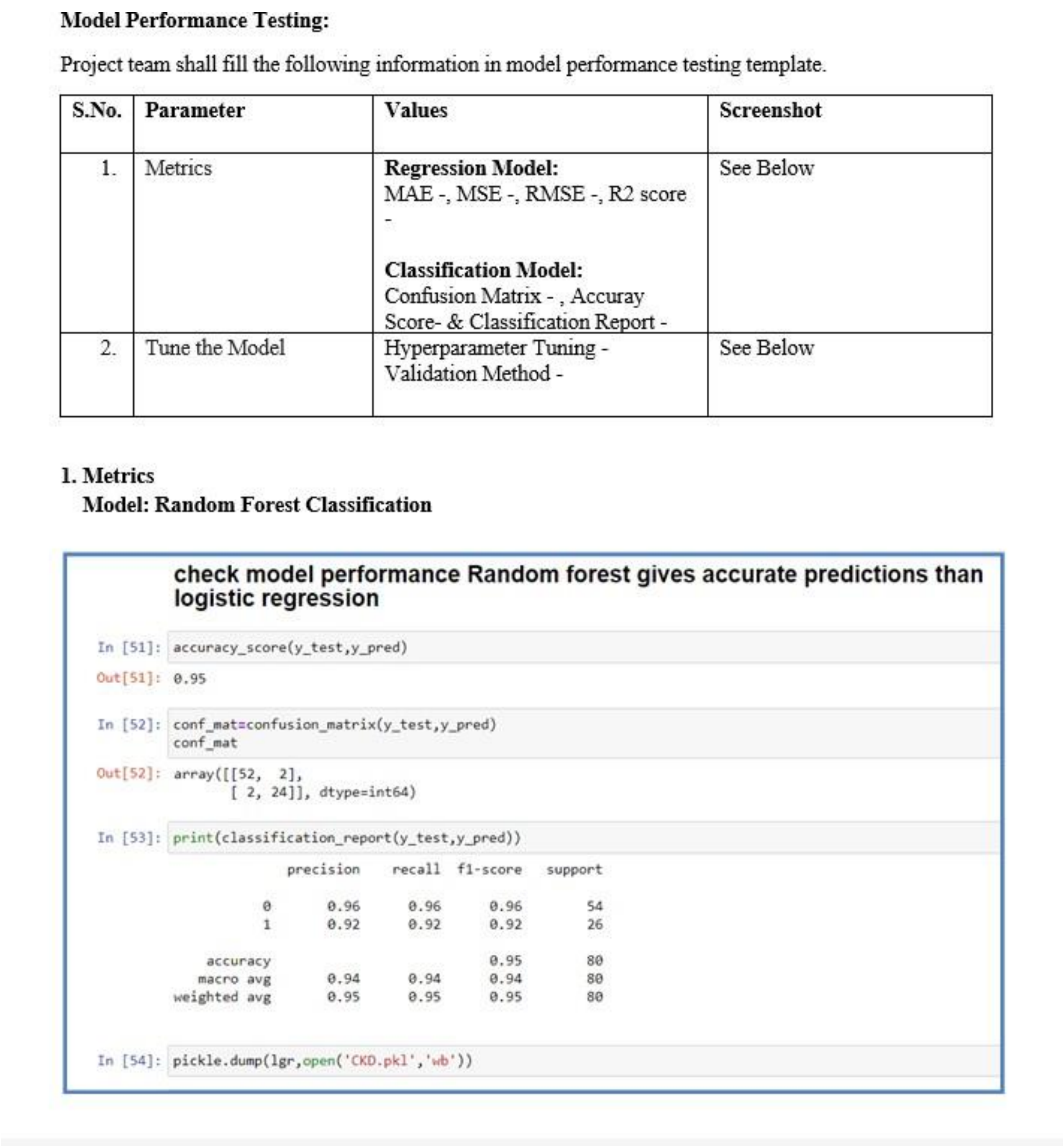




Test case ID	Feature Type	Component	Test Scenario	Pre-Requisite	Steps To Execute	Test Data	Expected Result	Actual Result	Status	Commnets	TC for Automation(Y/N)	BUG ID	Executed By
InitialScreen_TC_001	Functional	Home Page	Verify user able to see the Prediction page		1.Enter URL 2.Click on Prediction button 3.Verify going to next page		Entering into data input page	Working as expected	Pass	Normal test case			R.Sahana
Input_data_TC_002	Functional	Prediction value input page UI	Verify user able to enter input value		1.Check entering into prediction page 2.Check if user can enter value		Application should show below UI elements to enter numeric values: a.Blood Urea b.Blood Glucose Random Software should accept only numeric values	Should allow entering numeric values	Pass	Normal test case			S.Tamilarasi
Input_data_TC_003	Functional	Prediction value input page UI	Verify user able to enter input value		1.Check entering into prediction page 2.Check if user can select option from drop down box		Application should show below UI elements to select from drop down menu: a.Select Anemia b.Select Coronary Artery Disease c.Select Pus Cell d.Select Red Blood Cell e.Select Diabetics Mellitus	should allow selection from pull down menu	Pass	Normal test case			D.Kanimozhi
Input data_TC_004	Functional	Prediction value input page UI	Verify user able to enter input value		1.Check entering into prediction page 2.Check if user can select option from drop down box		Application should show below UI elements to enter alphabetic characters: a.Blood Urea b.Blood Glucose Random Software should accept only numeric values	Should not allow entering alphabetic values	Pass	Robustness test case			V.Keerthana
Result_data_TC_005	Functional	Prediction Result Page	Verify Chronic Kidney Disease (CKD) test values		1.Enter submit button after entering values 2.Redirect to result page and display correct result	a.Blood Urea : 90 b.Blood Glucose Random : 157 c.Select Anemia : No d.Select Coronary Artery Disease : Yes e.Select Pus Cell :Yes f.Select Red Blood Cell :No g.Select Diabetics Mellitus : Yes h.Select Pedal Edema: Yes	Application should show Chronic Kidney Disease	Showed CKD	Pass	Normal test case			G.Bhavani
Result_data_TC_006	Functional	Prediction Result Page	Verify No Chronic Kidney Disease (No CKD) test values		1.Enter submit button after entering values 2.Redirect to result page and display correct result	a.Blood Urea : 46 b.Blood Glucose Random : 117 c.Select Anemia : No d.Select Coronary Artery Disease : No e.Select Pus Cell :No f.Select Red Blood Cell :No g.Select Diabetics Mellitus : No h.Select Pedal Edema: No	Application should show No Chronic Kidney Disease	Showed No CKD	Pass	Normal test case			R.Sahana
Result_data_TC_007	Functional	Prediction Result Page	Verify Chronic Kidney Disease (CKD) test values		1.Enter submit button after entering values 2.Redirect to result page and display correct result	a.Blood Urea : 148 b.Blood Glucose Random : 173 c.Select Anemia : Yes d.Select Coronary Artery Disease : Yes e.Select Pus Cell :No f.Select Red Blood Cell :No g.Select Diabetics Mellitus : Yes h.Select Pedal Edema: Yes	Application should show Chronic Kidney Disease	Showed CKD	Pass	Normal test case			V.Keerthana
Result_data_TC_008	Functional	Prediction Result Page	Verify No Chronic Kidney Disease (No CKD) test values		1.Enter submit button after entering values 2.Redirect to result page and display correct result	a.Blood Urea : 24 b.Blood Glucose Random : 132 c.Select Anemia : No d.Select Coronary Artery Disease : No e.Select Pus Cell :No f.Select Red Blood Cell :No g.Select Diabetics Mellitus : No h.Select Pedal Edema: No	Application should show No Chronic Kidney Disease	Showed No CKD	Pass	Normal test case			D.Kanimozhi

9. RESULTS

9.1 Performance Metrics:



Model Performance Testing:

Project team shall fill the following information in model performance testing template.

S.No.	Parameter	Values	Screenshot
1.	Metrics	**Regression Model:** MAE -, MSE -, RMSE -, R2 score - **Classification Model:** Confusion Matrix - , Accuray Score- & Classification Report -	See Below
2.	Tune the Model	Hyperparameter Tuning - Validation Method -	See Below

1. Metrics

Model: Random Forest Classification

check model performance Random forest gives accurate predictions than logistic regression

```
In [51]: accuracy_score(y_test,y_pred)

Out[51]: 0.95

In [52]: conf_mat=confusion_matrix(y_test,y_pred)
         conf_mat

Out[52]: array([[52,  2],
                [ 2, 24]], dtype=int64)

In [53]: print(classification_report(y_test,y_pred))

                       precision    recall  f1-score   support

                    0       0.96      0.96      0.96        54
                    1       0.92      0.92      0.92        26

             accuracy                           0.95        80
            macro avg       0.94      0.94      0.94        80
         weighted avg       0.95      0.95      0.95        80


In [54]: pickle.dump(lgr,open('CKD.pkl','wb'))
```

2. Tune the Model

Hyperparameter Tuning:

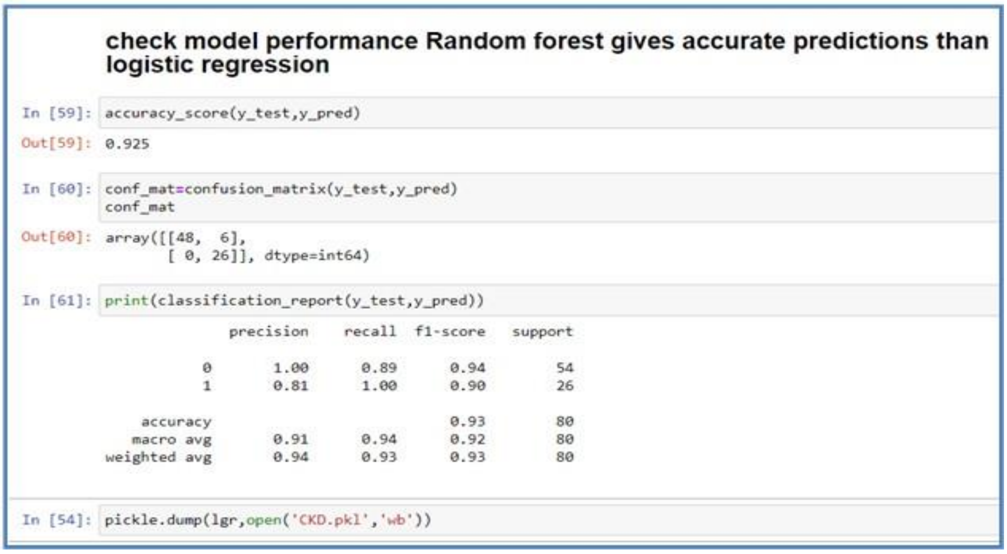
- The number of features is important and should be tuned in random forest classification.
- Initially all parameters in the dataset are taken as independent values to arrive at the dependent decision of Chronic Kidney Disease or No Chronic Kidney Disease.
- But the result was not accurate so used only 8 more correlated values as independent values to arrive at the dependent decision of Chronic Kidney Disease or not.

Validation Method:

It involves **partitioning the training data set into subsets, where one subset is held out to test the performance of the model**. This data set is called the validation data set.

Cross validation is to use different models and identify the best:

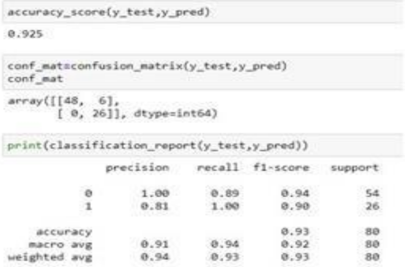
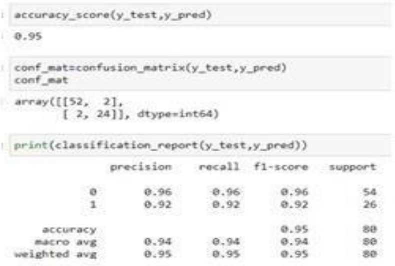
Logistic Regression Model performance values:



check model performance Random forest gives accurate predictions than logistic regression

```
In [59]: accuracy_score(y_test,y_pred)

Out[59]: 0.925

In [60]: conf_mat=confusion_matrix(y_test,y_pred)
         conf_mat

Out[60]: array([[48,  6],
                [ 0, 26]], dtype=int64)

In [61]: print(classification_report(y_test,y_pred))

                       precision    recall  f1-score   support

                    0       1.00      0.89      0.94        54
                    1       0.81      1.00      0.90        26

             accuracy                           0.93        80
            macro avg       0.91      0.94      0.92        80
         weighted avg       0.94      0.93      0.93        80

In [54]: pickle.dump(lgr,open('CKD.pkl','wb'))
```

Hence we tested with Logistic regression and Random Forest Classification wherein the accuracy of Random Forest classification is 99% compared with Logistic Regression.

Metric	Logistic Regression	Random Forest Classification
Accuracy	0.925	0.95
Other metrics	 accuracy_score(y_test,y_pred) 0.925 conf_mat=confusion_matrix(y_test,y_pred) conf_mat array([[48, 6], [0, 26]], dtype=int64) print(classification_report(y_test,y_pred)) precision recall f1-score support 0 1.00 0.89 0.94 54 1 0.81 1.00 0.90 26 accuracy 0.93 80 macro avg 0.91 0.94 0.92 80 weighted avg 0.94 0.93 0.93 80	 accuracy_score(y_test,y_pred) 0.95 conf_mat=confusion_matrix(y_test,y_pred) conf_mat array([[52, 2], [2, 24]], dtype=int64) print(classification_report(y_test,y_pred)) precision recall f1-score support 0 0.96 0.96 0.96 54 1 0.92 0.92 0.92 26 accuracy 0.95 80 macro avg 0.94 0.94 0.94 80 weighted avg 0.95 0.95 0.95 80

The above table shows that Random Forest Classification gives better results over Logistic Regression.

10. ADVANTAGES & DISADVANTAGES

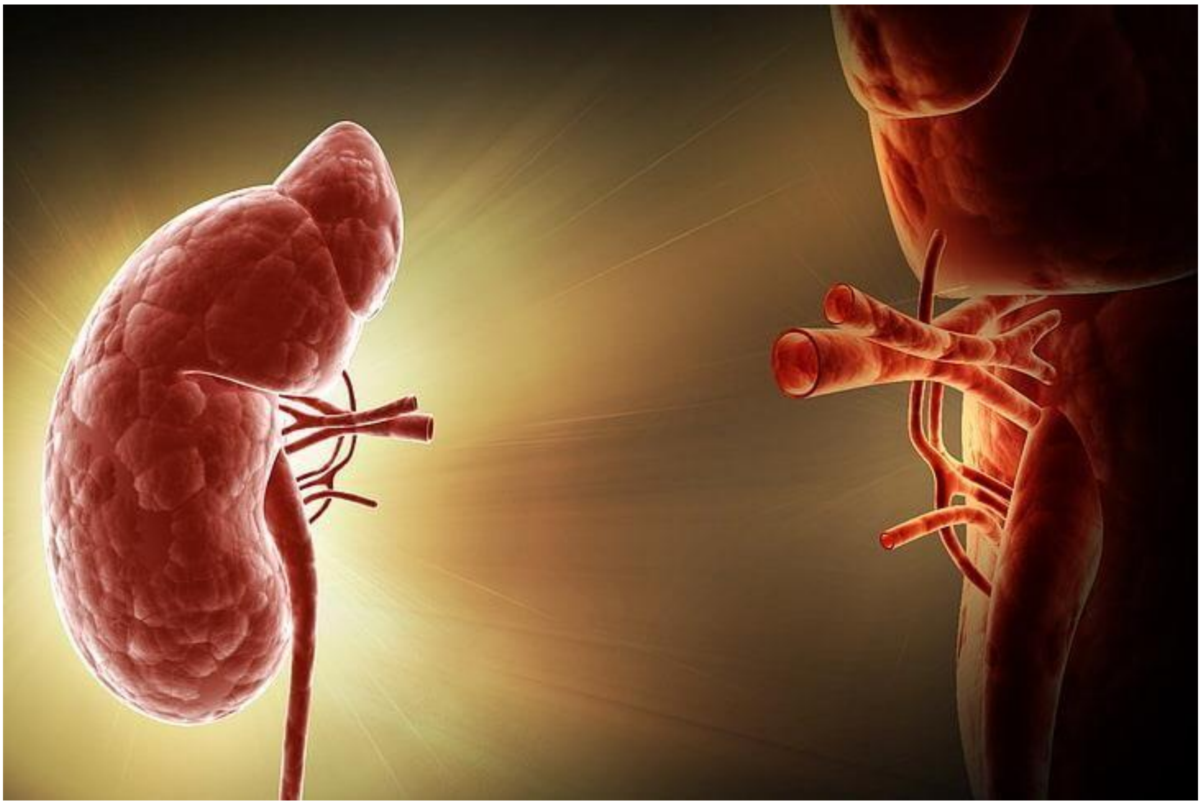
- This software has various advantages where it can be used as an expert guide to doctors for early detection of chronic kidney disease. It is also seen in performance metrics that it has an accuracy of 95% which gives good confidence to the users.

- Machine Learning is autonomous but highly susceptible to errors. Suppose you train an algorithm with data sets small enough to not be inclusive. You end up with biased predictions coming from a biased training set.

11. CONCLUSION

- This Project has helped team members to understand various concepts of Machine learning, Flask file, IBM cloud and Python notebook.

- This project can be scaled for usage in prediction of other chronic diseases which will help doctors in diagnosis of disease at an early stage thereby helping in early detection of various disease.



12. FUTURE SCOPE

This software can be used to detect various other chronic diseases by modifying the dataset and the user inputs received. The model can be further trained with enormous amount of data to improve the accuracy.

13. APPENDIX

13.1 Source Code:

https://github.com/IBM-EPBL/IBM-Project-433-1658301077/tree/main/Project%20Development%20Phase

13.2 GitHub & Project Demo Link:

Github: **https://github.com/IBM-EPBL/IBM-Project-433-1658301077/tree/main/Project%20Development%20Phase**

Demo Link: https://youtu.be/1csfumlsdYE