**Early Prediction For Chronic Kidney Disease Detection: A Progressive Approach To Health Management**

* **divya.p**
* **gayathri devi. s**
* **haripriya. m**
* **hemabharathi. v**
* **thasleema. j**

|  |  |
| --- | --- |
| **S. NO** | **CONTENT** |
| 1 | INTRODUCTION |
| 2 | PROBLEM DEFINITION & DESIGN THINKING |
| 3 | RESULT |
| 4 | ADVANTAGES & DISADVANTAGES |
| 5 | APPLICATIONS |
| 6 | CONCLUSION |
| 7 | FUTURE SCOPE |
| 8 | APPENDIX |

**INTRODUCTION**

**Early Prediction For Chronic Kidney Disease Detection: A Progressive Approach To Health Management**

### Define Problem / Problem Understanding

### Data Collection & Preparation

### Exploratory Data Analysis

### Model Building

### Performance Testing & Evaluate The Results

### Model Deployment

### Project Demonstration & Documentation

**Early Prediction For Chronic Kidney Disease Detection: A Progressive Approach To Health Management**

**Project Description:**

Chronic Kidney Disease (CKD) is a major medical problem and can be cured if treated in the early stages. Usually, people are not aware that medical tests we take for different purposes could contain valuable information concerning kidney diseases. Consequently, attributes of various medical tests are investigated to distinguish which attributes may contain helpful information about the disease. The information says that it helps us to measure the severity of the problem, the predicted survival of the patient after the illness, the pattern of the disease and work for curing the disease.

In todays world as we know most of the people are facing so many disease and  as this can be cured if we treat people in early stages this project can use a pretrained model to predict the Chronic Kidney Disease which can help in treatments of peoples who are suffer from this disease.

**Technical Architecture:**



**Project Flow:**

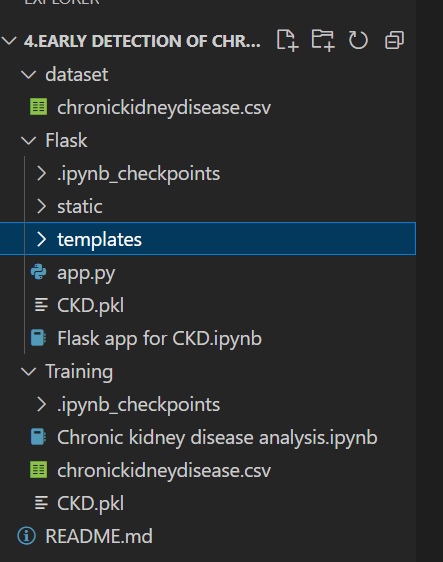
* User interacts with the UI to enter the input.
* Entered input is analysed by the model which is integrated.
* Once model analyses the input the prediction is showcased on the UI

To accomplish this, we have to complete all the activities listed below,

* Define Problem / Problem Understanding
  + Specify the business problem
  + Business requirements
  + Literature Survey
  + Social or Business Impact.
* Data Collection & Preparation
  + Collect the dataset
  + Data Preparation
* Exploratory Data Analysis
  + Descriptive statistical
  + Visual Analysis
* Model Building
  + Training the model in multiple algorithms
  + Testing the model
* Performance Testing & Evaluate the results
  + Testing model with multiple evaluation metrics
  + Evaluate the results
* Model Deployment
  + Save the best model
  + Integrate with Web Framework
* Project Demonstration & Documentation
  + Record explanation Video for project end to end solution
  + Project Documentation-Step by step project development procedure

**Project Structure:**

Create the Project folder which contains files as shown below



* We are building a flask application which needs HTML pages stored in the templates folder and a python script app.py for scripting.
* CKD.pkl is our saved model. Further we will use this model for flask integration.
* Training folder contains a model training file.

### 1.1 OVERVIEW

### Define Problem / Problem Understanding

* In This milestone, we will see the problem understanding

### Specify The Business Problem

* Chronic Kidney Disease (CKD) is a major medical problem and can be cured if treated in the early stages. Usually, people are not aware that medical tests we take for different purposes could contain valuable information concerning kidney diseases. Consequently, attributes of various medical tests are investigated to distinguish which attributes may contain helpful information about the disease. The information says that it helps us to measure the severity of the problem, the predicted survival of the patient after the illness, the pattern of the disease and work for curing the disease.
* In todays world as we know most of the people are facing so many disease and  as this can be cured if we treat people in early stages this project can use a pretrained model to predict the Chronic Kidney Disease which can help in treatments of peoples who are suffer from this disease.

### 1.2 PURPOSE

### Business Requirements

The business requirements for a machine learning model to predict chronic kidney disease include the ability to accurately predict the ckd based on given information, Minimise the number of false positives (predicting diseased) and false negatives (not diseased). Provide an explanation for the model's decision, to comply with regulations and improve transparency.

### Literature Survey

Chronic kidney disease (CKD) is a significant public health issue, affecting an estimated 14% of the global population. The disease is characterized by a gradual loss of kidney function over time, leading to a range of serious health complications, including end-stage renal disease (ESRD) requiring dialysis or kidney transplant. Early detection and management of CKD is crucial to prevent progression to ESRD and improve patient outcomes.

There have been numerous studies in recent years aimed at developing accurate and efficient methods for predicting CKD progression. These studies have employed a variety of techniques, including machine learning, deep learning, and artificial neural networks.

### Social Or Business Impact

On a social level, early detection and prediction of CKD can lead to improved patient outcomes and quality of life. By identifying individuals at risk for CKD, healthcare providers can intervene early and slow the progression of the disease through lifestyle changes, medication management, and other treatments. This can help prevent the need for dialysis or kidney transplantation, which can be costly and life-altering for patients. Additionally, early prediction can also help reduce the overall burden of CKD on the healthcare system by reducing the number of hospitalizations and emergency room visits.

### Data Collection & Preparation

ML depends heavily on data. It is the most crucial aspect that makes algorithm training possible. So this section allows you to download the required dataset.

### Collect The Dataset

There are many popular open sources for collecting the data. Eg: kaggle.com, UCI repository, etc.

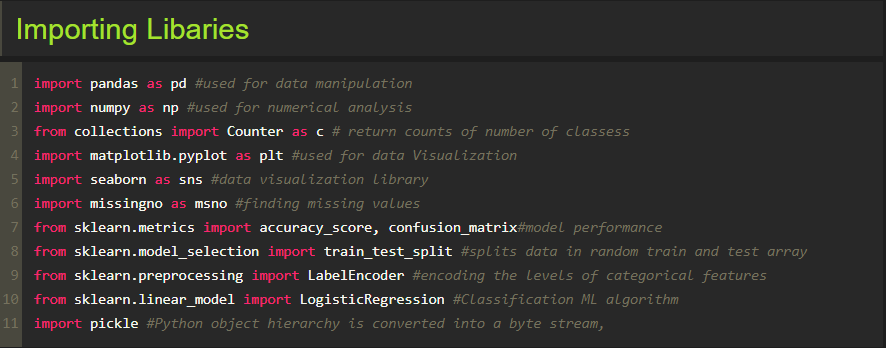
In this project we have used .csv data. This data is downloaded from kaggle.com. Please refer to the link given below to download the dataset

As the dataset is downloaded. Let us read and understand the data properly with the help of some visualisation techniques and some analysing techniques.

**Note:** There are several techniques for understanding the data. But here we have used some of it. In an additional way, you can use multiple techniques.

**Importing The Libraries**

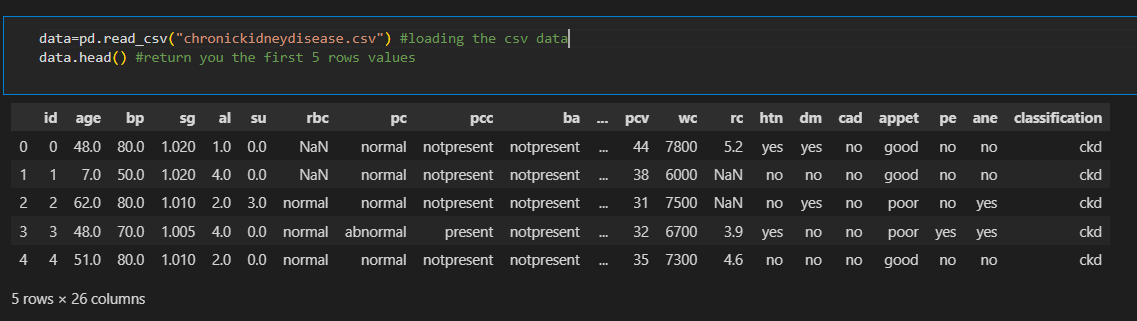
Import the necessary libraries as shown in the image. (optional) Here we have used visualisation style as fivethirtyeight.



**Read The Dataset**

Our dataset format might be in .csv, excel files, .txt, .json, etc. We can read the dataset with the help of pandas.

In pandas we have a function called read\_csv() to read the dataset. As a parameter we have to give the directory of the csv file.



### Data Preparation

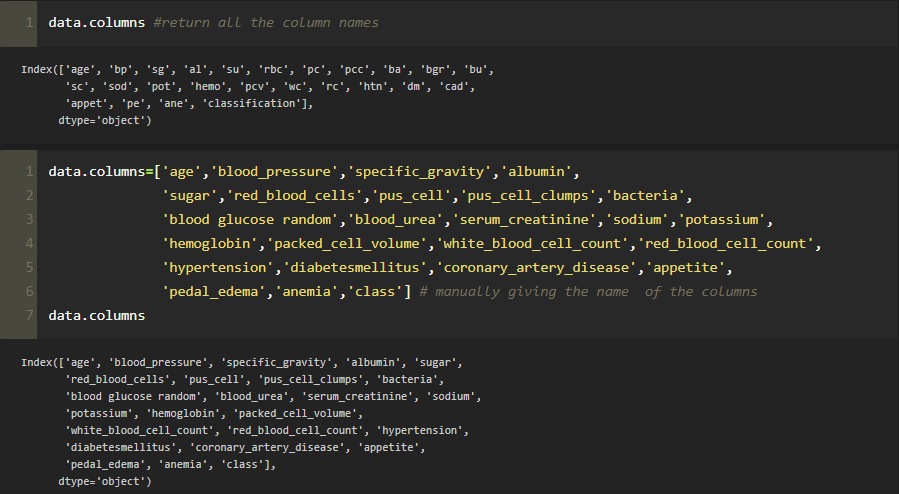
As we have understood how the data is, let's pre-process the collected data.

The download data set is not suitable for training the machine learning model as it might have so much randomness so we need to clean the dataset properly in order to fetch good results. This activity includes the following steps.

* Rename the columns
* Handling missing values
* Handling categorical data
* Handling Numerical data

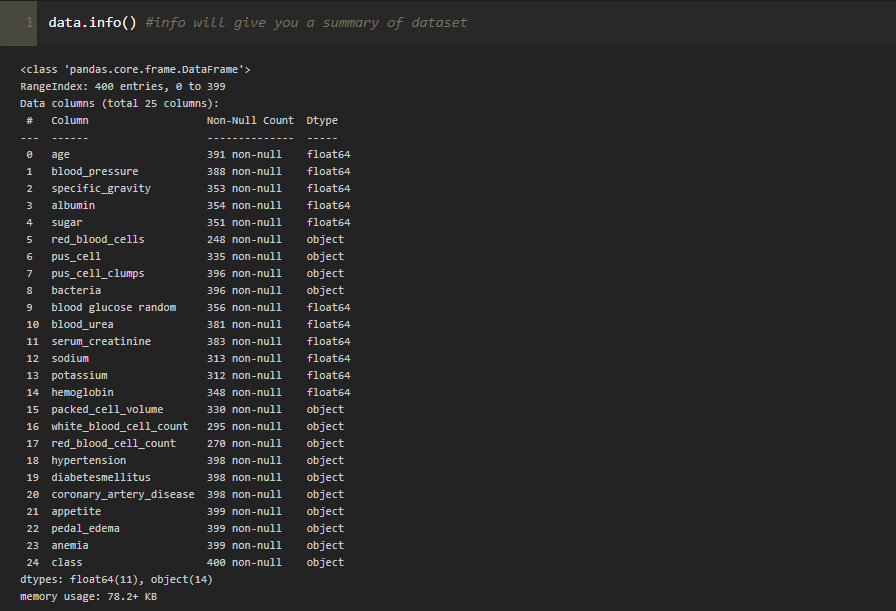
Note: These are the general steps of pre-processing the data before using it for machine learning. Depending on the condition of your dataset, you may or may not have to go through all these steps.

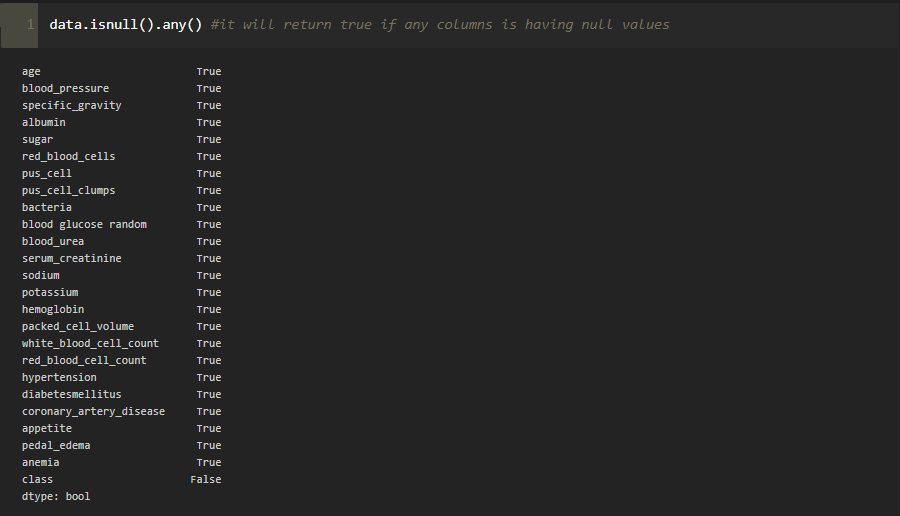
**Rename The Columns**

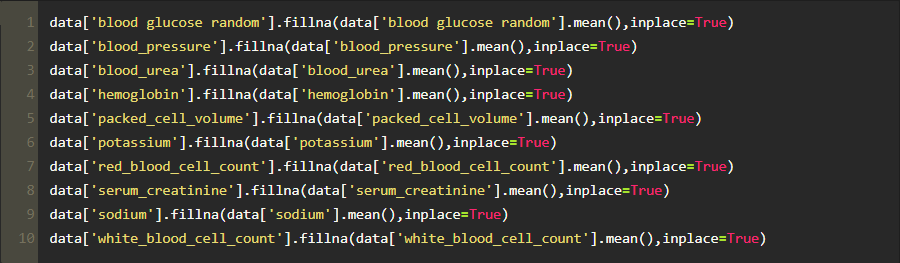


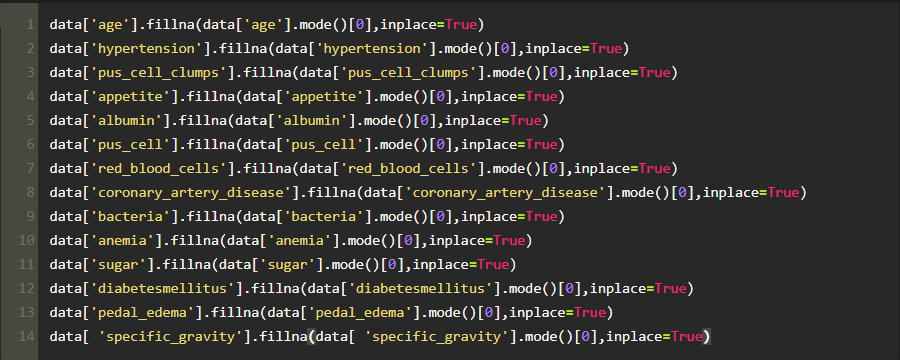
### Handling Missing Values

* Let’s find the shape of our dataset first. To find the shape of our data, the df.shape method is used. To find the data type, df.info() function is used.





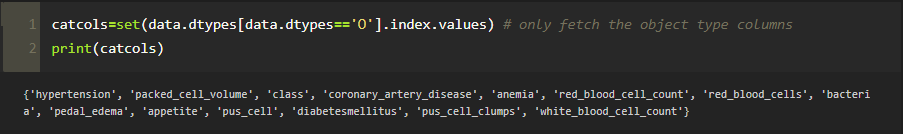




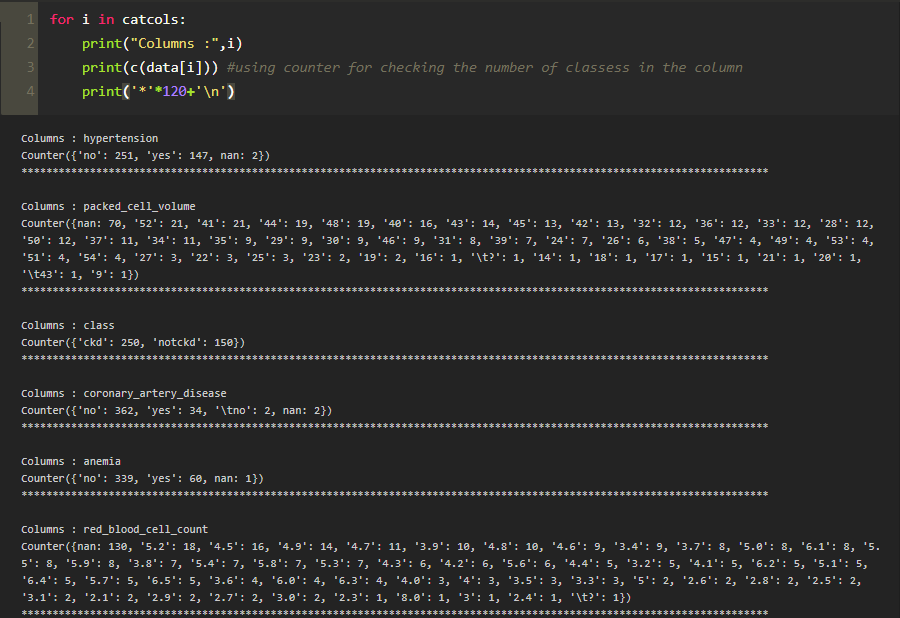
Let’s now check the count of null values after filling all null values using **isnull.sum()**

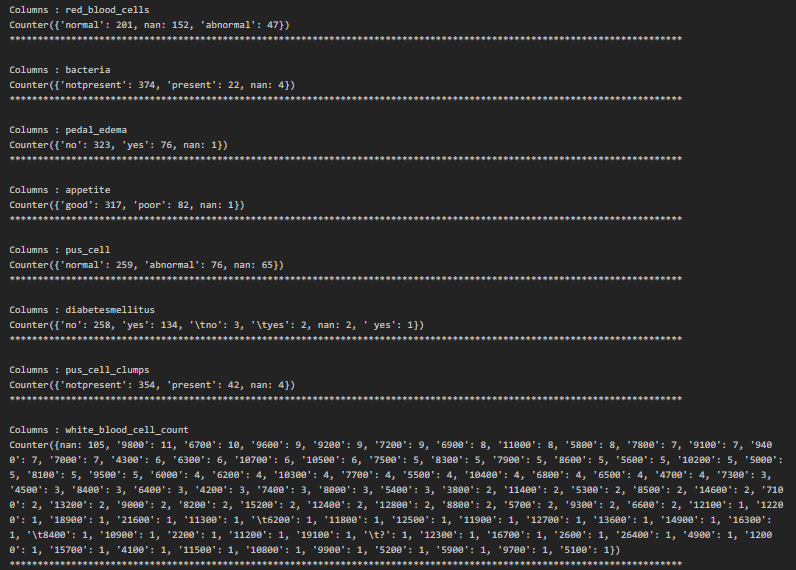
**Handling Categorical Columns**

The below code is used for fetching all the object or categorical type of columns from our data and we are storing it as **set** in variable **catcols.**



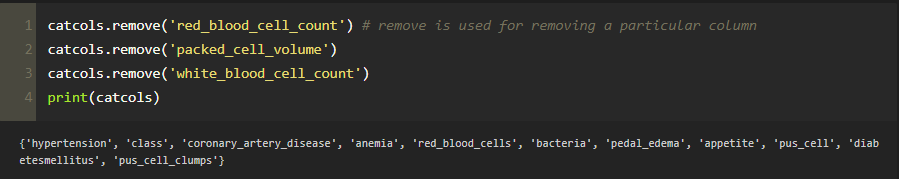
As, you can observe that it gives us the same count of columns which we find previously.





In the above we are looping with each categorical column and printing the classes of each categorical columns using counter function so that we can detect which columns are categorical and which are not.

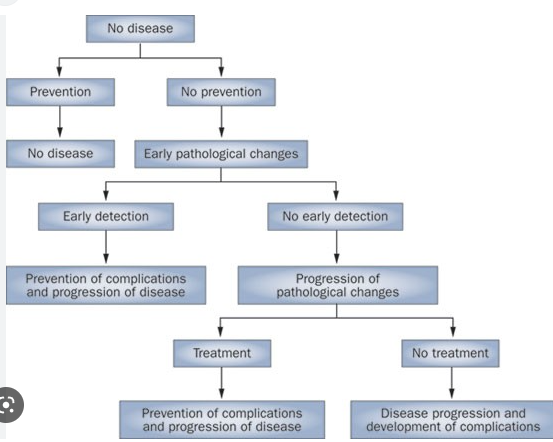
If you observe some columns have a few classes and some have many, those columns are having many classes can be considered as numerical column and we have to remove it and add it to the continuous columns.



As we store our columns as set, we can make use of **remove** function which is used to remove the element in our case we can take it as columns.

**2.PROBLEM DEFINITION & DESIGN THINKING**

**2.1 EMPATHY MAP**



**Label Encoding For Categorical Columns**

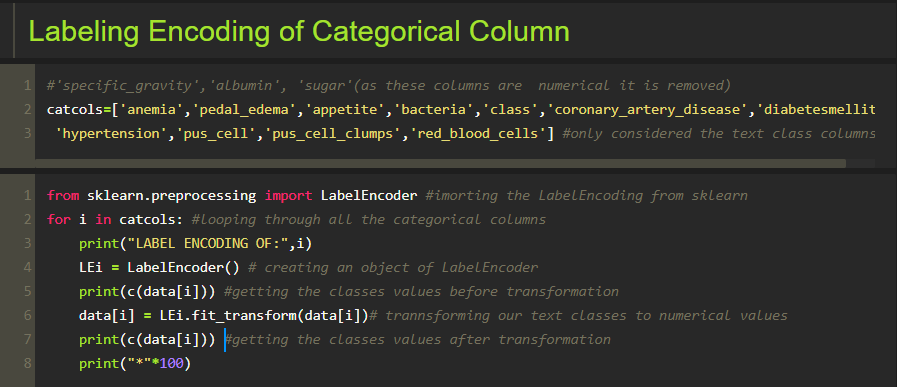
Typically, any structured dataset includes multiple columns with combination of numerical as well as categorical variables. A machine can only understand the numbers. It cannot understand the text. That’s essentially the case with [Machine Learning algorithms](https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/?utm_source=blog&utm_medium=one-hot-encoding-vs-label-encoding-using-scikit-learn) too.We need to convert each text category to numbers in order for the machine to process those using mathematical equations.

How should we handle categorical variables? There are Multiple way to handle, but will see one of it is LabelEncoding.

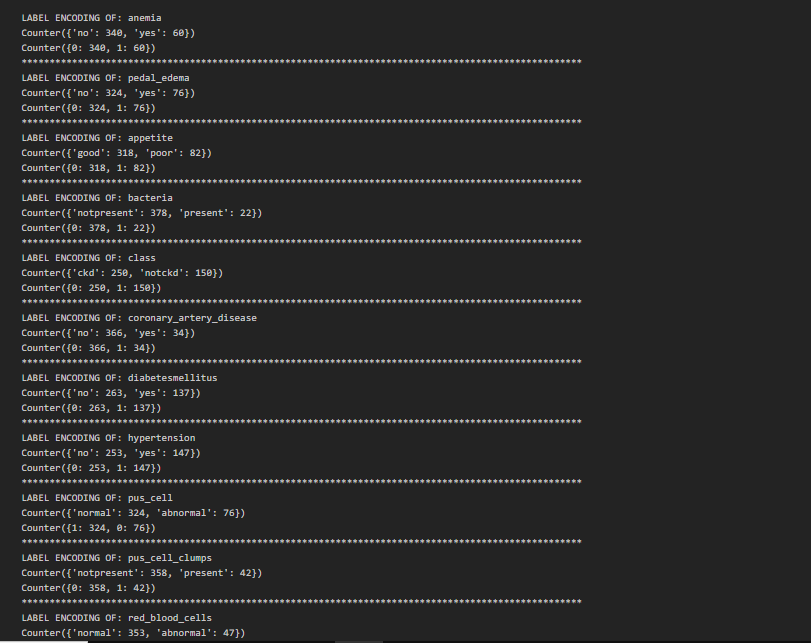
**Label Encoding** is a popular encoding technique for handling categorical variables. In this technique, each label is assigned a unique integer based on alphabetical ordering.

Let’s see how to implement label encoding in Python using the [scikit-learn library](https://www.analyticsvidhya.com/blog/2020/02/everything-you-should-know-scikit-learn/?utm_source=blog&utm_medium=one-hot-encoding-vs-label-encoding-using-scikit-learn).

we have to convert only the text class category columns; we first select it then we will implement Label Encoding to it.

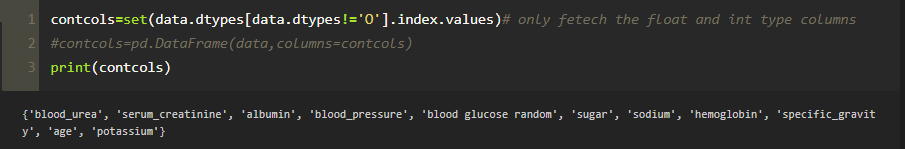


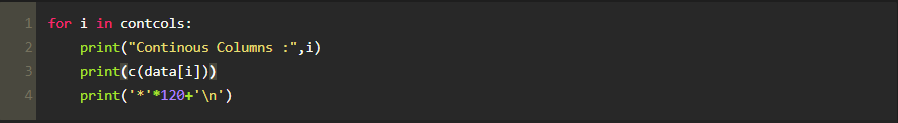
In the above code we are looping through all the selected text class categorical columns and performing label encoding.



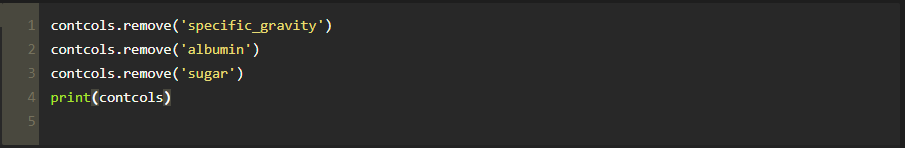
As you can see here, after performing label encoding alphabetical classes is converted to numeric.

**Handling Numerical Columns**

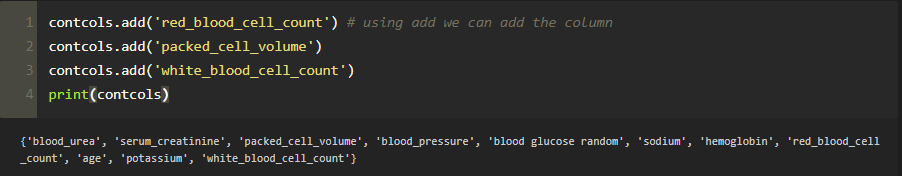
Same as we did with categorical columns, we are majing use of **dtypes** for finding the continuous columns

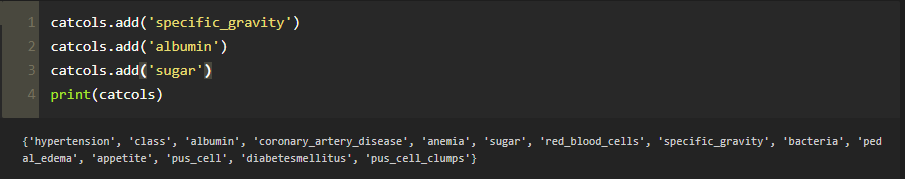


If we observe the output of the above code we can observe that some columns have few values or you can say classes which can be considered as categorical columns. So , let’s remove it and add the columns which we observed into their respective variables.

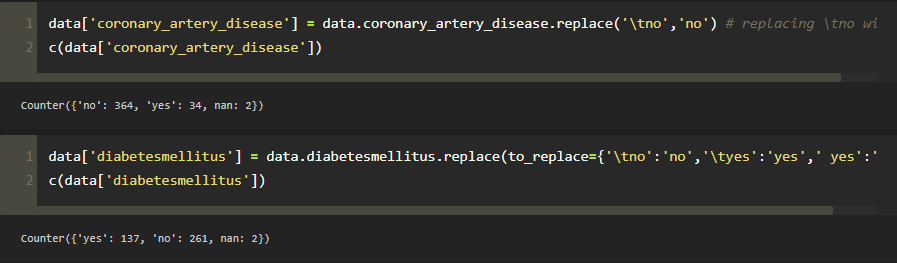


With the help of add() function we can add an element.





In our data some columns some unwanted classes so we have to rectify that also for that we simply use **replace()**



**2.2 IDEATION & DESIGN THINKING**

**Exploratory Data Analysis**

In this milestone, we will see the exploratory data analysis.

**Descriptive Statistical Analysis**

Descriptive analysis is to study the basic features of data with the statistical process. Here pandas has a worthy function called describe. With this describe function we can understand the unique, top and frequent values of categorical features. And we can find mean, std, min, max and percentile values of continuous features.

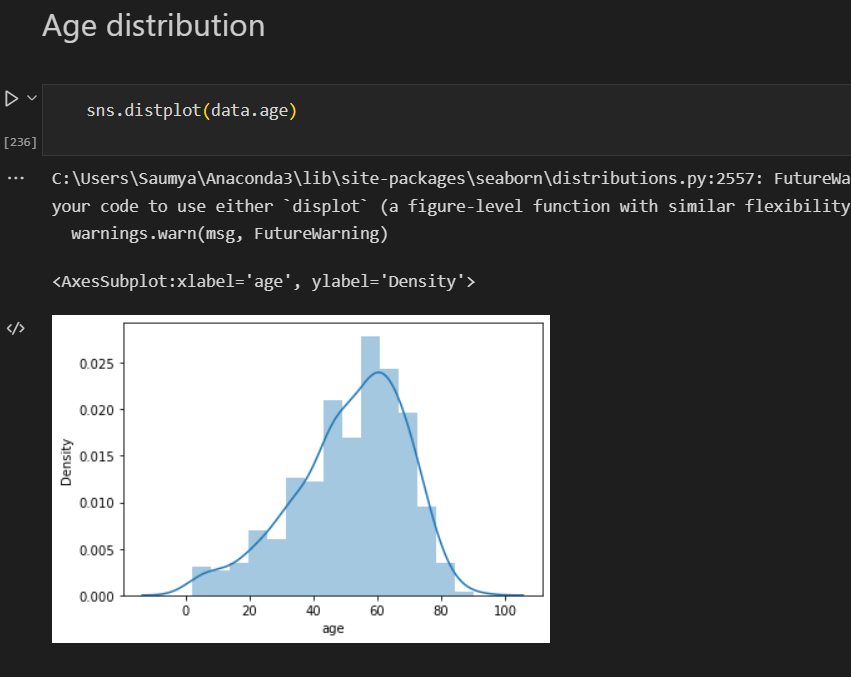
### Visual Analysis

Visual analysis is the process of using visual representations, such as charts, plots, and graphs, to explore and understand data. It is a way to quickly identify patterns, trends, and outliers in the data, which can help to gain insights and make informed decisions.

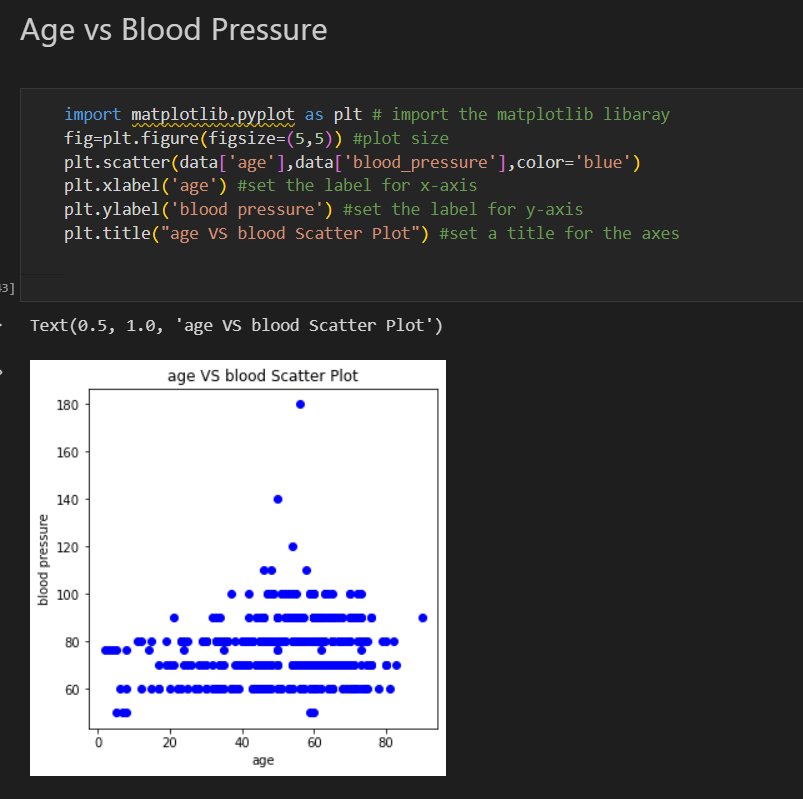
**Univariate Analysis**

In simple words, univariate analysis is understanding the data with a single feature. Here we have displayed two different graphs such as distplot and countplot.

The Seaborn package provides a wonderful function distplot. With the help of distplot, we can find the distribution of the feature.

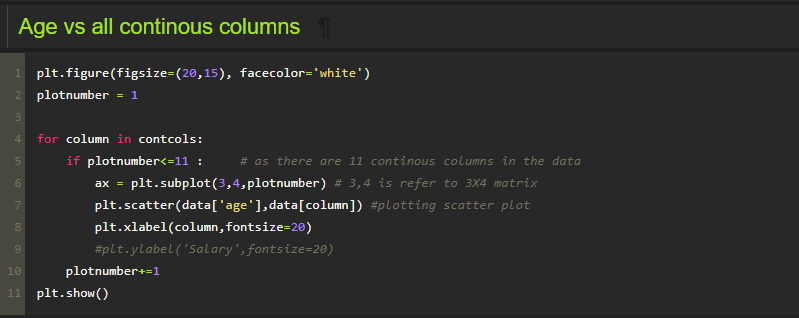


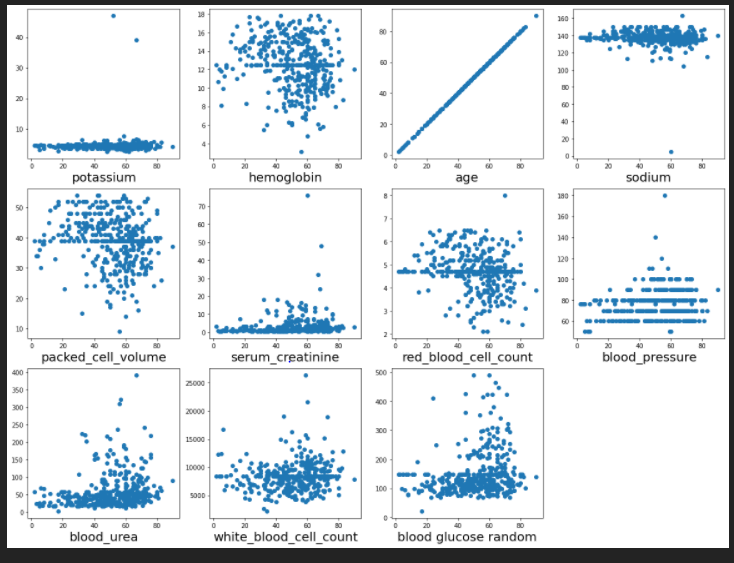
**Bivariate Analysis**

****

**Multivariate Analysis**

**Age vs all continuous columns**



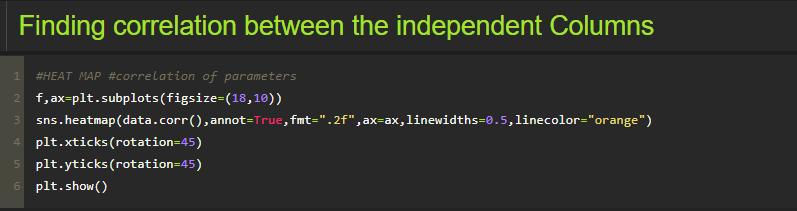


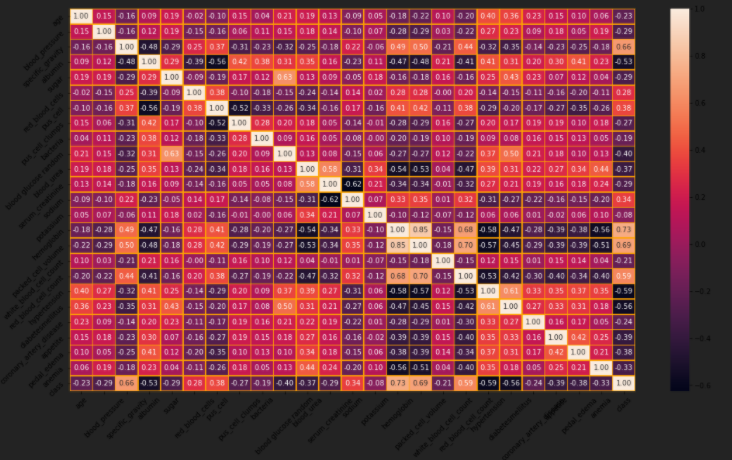
As you can observe with the scatter plot many of features are correlated with age.

**Finding correlation between the independent Columns**

Correlation is a statistical relationship between two variables and it could be positive, meaning both variables move in the same direction, or negative, meaning that when one variable's value increases, the other variables' values decrease.

With the help of seaborn heatmap we will be plotting the heatmap and for finding the correlation between variable we have **corr()** available.





If you observe the heatmap, lighter the colour the correlation between that two variables will be high.

And correlation plays a very important role for extracting the correct features for build our model

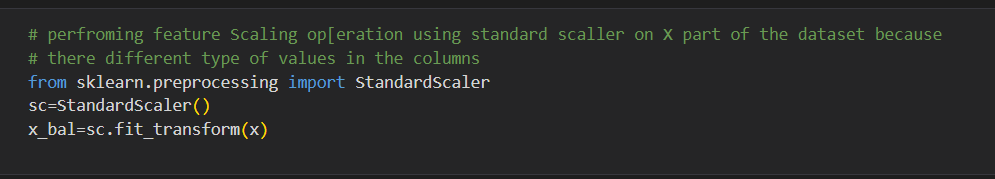
Now, let’s observe the count of our target data classes, by using seaborn countplot



**Scaling the Data**

Scaling is one the important process, we have to perform on the dataset, because of data measures in different ranges can leads to mislead in prediction

Models such as KNN, Logistic regression need scaled data, as they follow distance-based method and Gradient Descent concept.



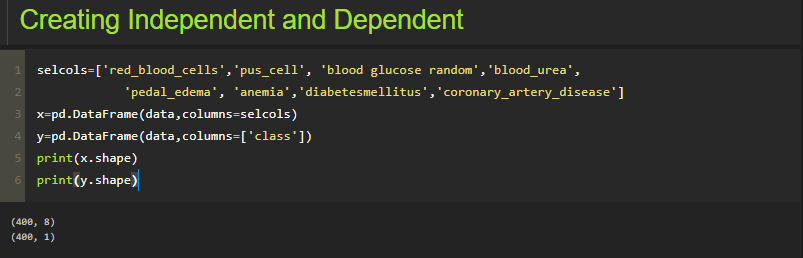
We will perform scaling only on the input values.Once the dataset is scaled, it will be converted into an array and we need to convert it back to a dataframe.

**Separate independent and dependent variable**

Now let’s split the Dataset into train and test sets

Changes: first split the dataset into x and y and then split the data set

Here x and y variables are created. On x variable, df is passed with dropping the target variable. And on y target variable is passed. For splitting training and testing data we are using the train\_test\_split() function from sklearn. As parameters, we are passing x, y, test\_size, random\_state.

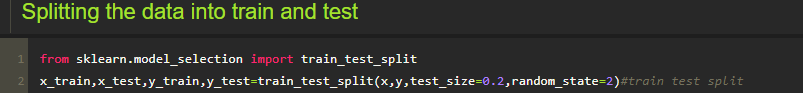


In the above code we are creating DataFrame of the independent variable **x** with our selected columns and for dependent variable **y** we are only taking the **class** column.

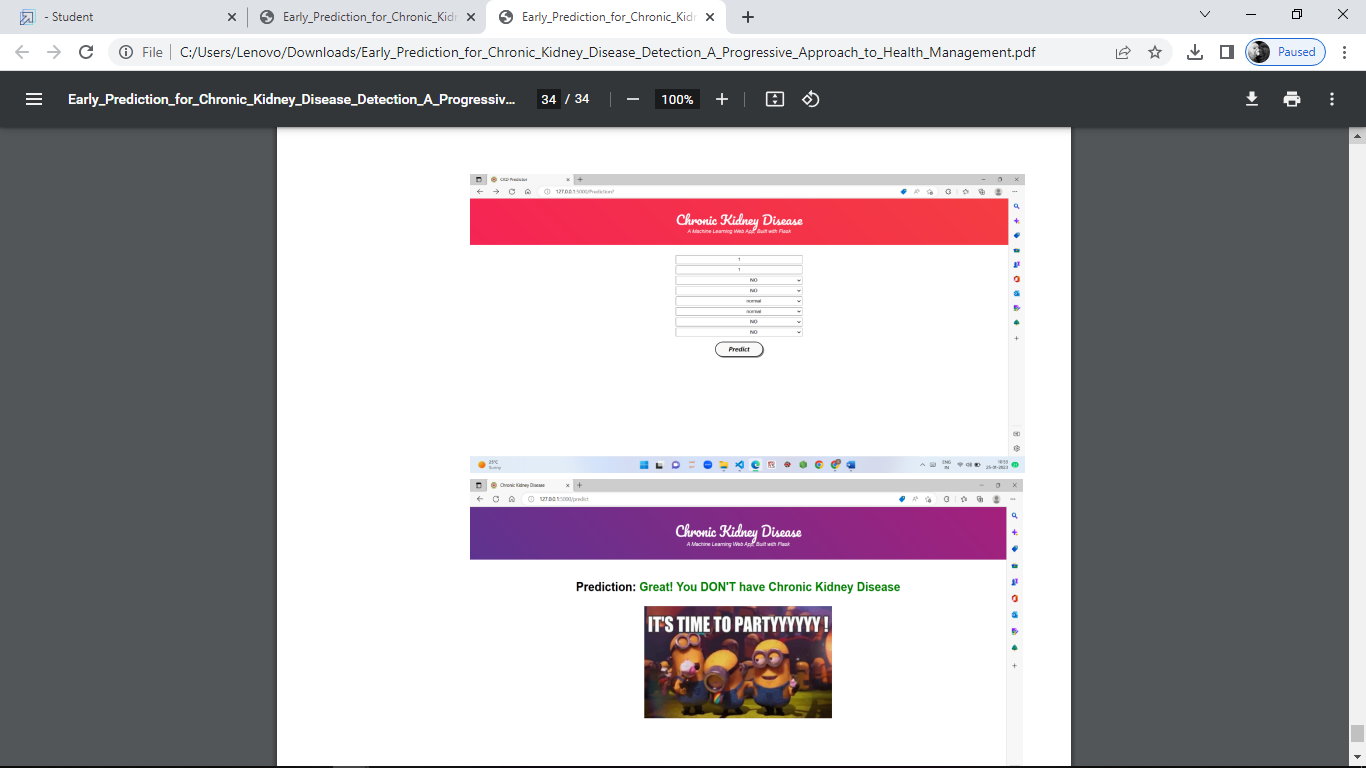
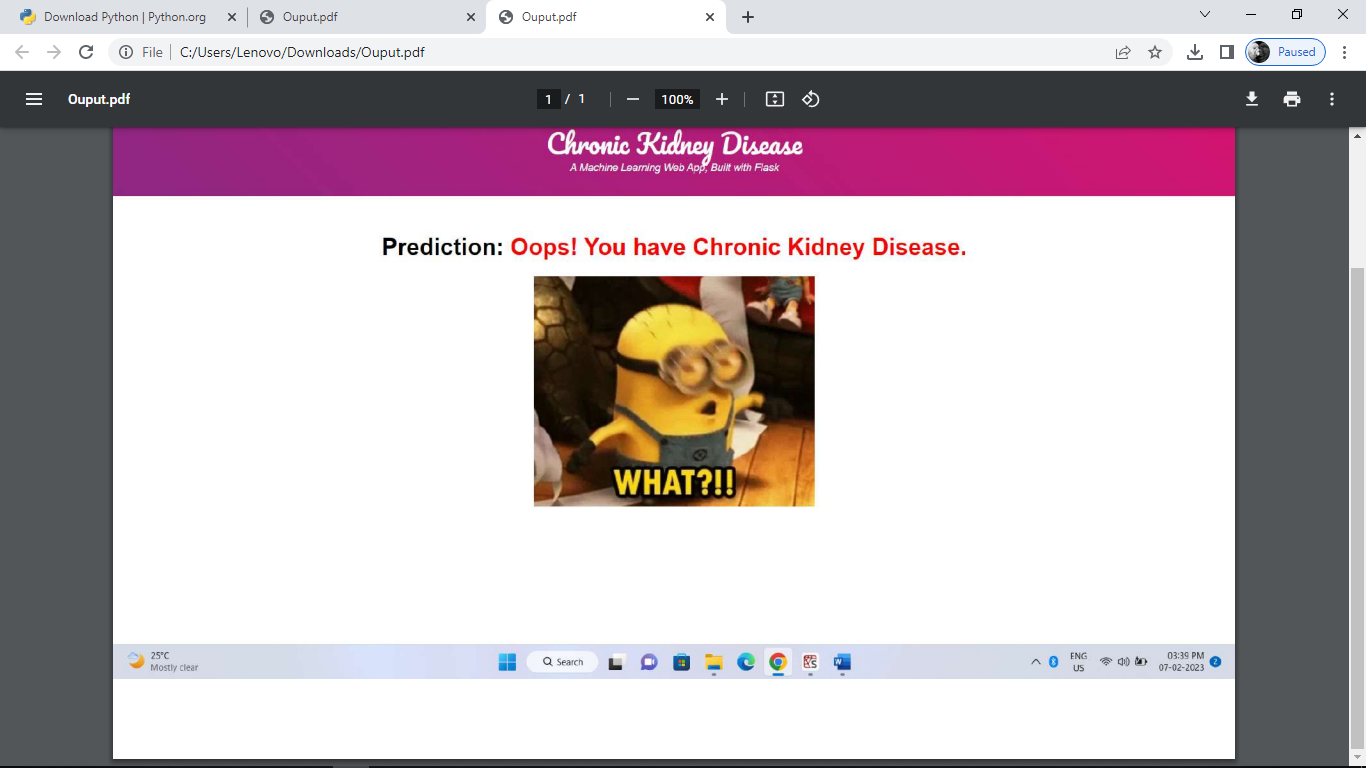
Where **DataFrame is used to** represents a table of data with rows and columns.

**Splitting data into train and test**

When you are working on a model and you want to train it, you obviously have a dataset. But after training, we have to test the model on some test dataset. For this, you will a dataset which is different from the training set you used earlier. But it might not always be possible to have so much data during the development phase. In such cases, the solution is to split the dataset into two sets, one for training and the other for testing.



**3. RESULT**



**4. ADVANTAGES & DISADVANTAGES**

**ADVANTAGES**

One of the main advantages of PD over hemodialysis is that the procedure can be carried out in the comfort of the patients’ home. For most, all that is required is a washroom with fresh running water, a sterile area of the house for the procedure to take place, and space to store the fluid for dialysis. This also allows patients to travel.

For elderly patients who may be unable to administer the procedure themselves, assistance may be given by a trained carer or community nurse. Additionally, the procedure can even be run while the patient sleeps.

### 

**DISADVANTAGES**

One of the most significant disadvantages of PD is that it’s required to be carried out every day, which may act as an inconvenience for some. Furthermore, by undergoing PD specifically, there may be a greater risk of developing peritonitis – an infection of the abdomen.

In some, quite rare instances, patients may experience scarring and thickening of the peritoneum. In order to prevent this, patients may have to change their treatment to hemodialysis.

Additionally, those who undergo PD may report increased fatigue and malnutrition caused by the dialysis fluid, which can lower protein levels.

**5. APPLICATION**

**Model Building**

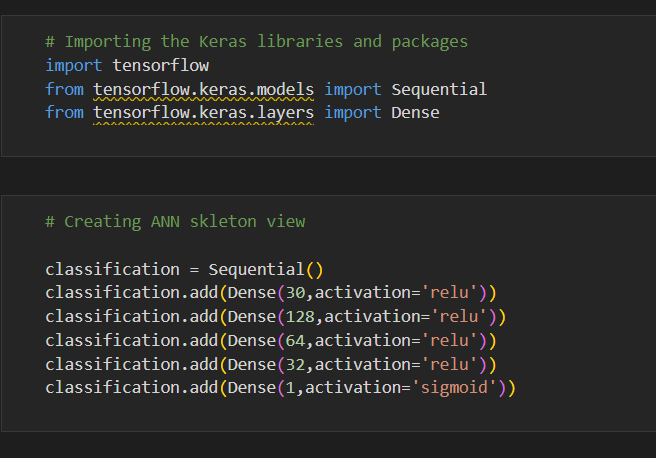
In this milestone, we will see the model building

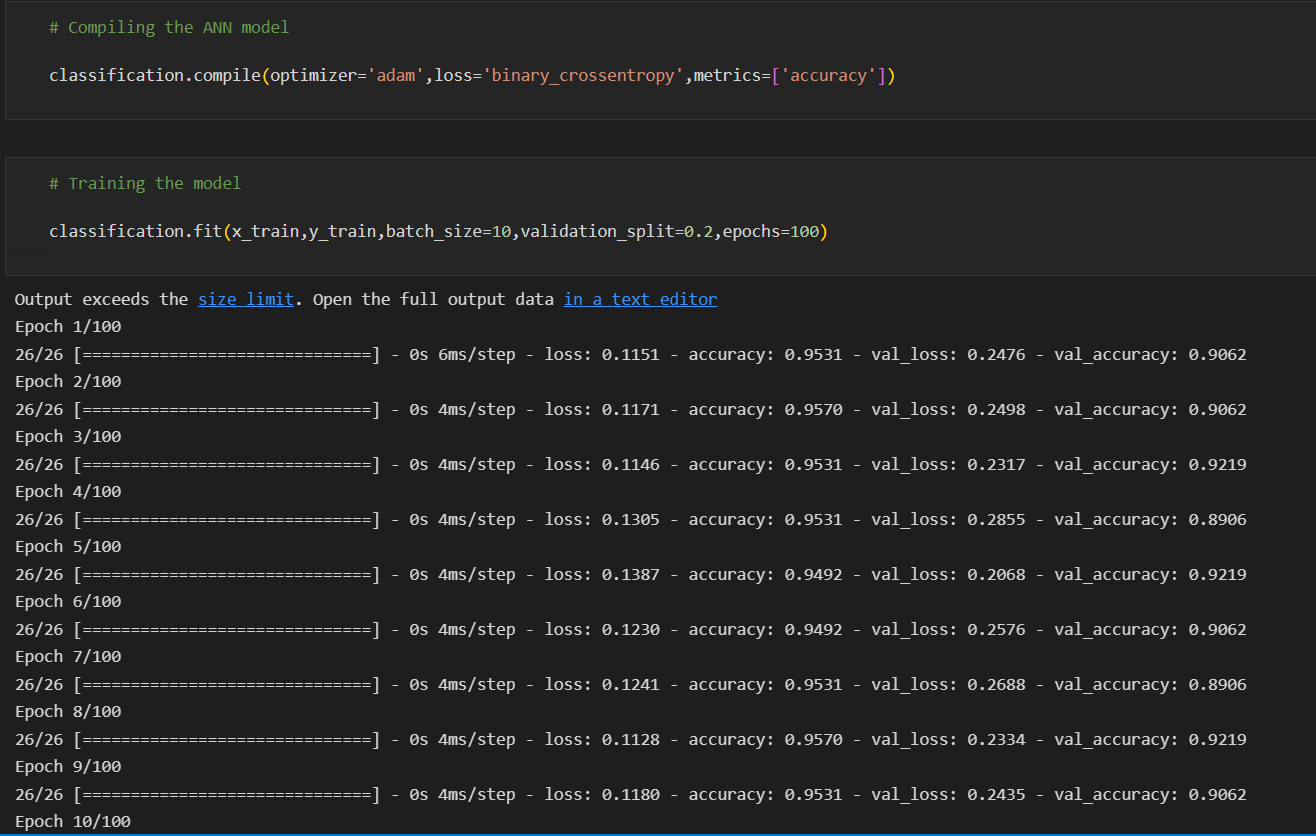
**Training The Model In Multiple Algorithms**

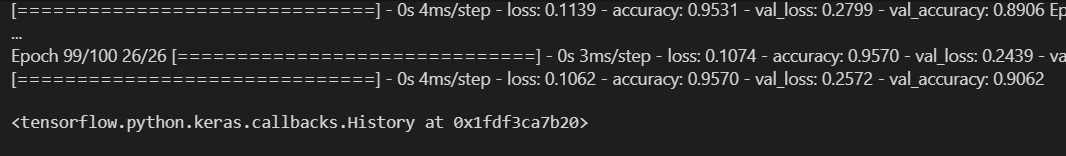
Now our data is cleaned and it’s time to build the model. We can train our data on different algorithms. For this project we are applying four classification algorithms. The best model is saved based on its performance.

**ANN Model**

Building and training an Artificial Neural Network (ANN) using the Keras library with TensorFlow as the backend. The ANN is initialised as an instance of the Sequential class, which is a linear stack of layers. Then, the input layer and two hidden layers are added to the model using the Dense class, where the number of units and activation function are specified. The output layer is also added using the Dense class with a sigmoid activation function. The model is then compiled with the Adam optimizer, binary cross-entropy loss function, and accuracy metric. Finally, the model is fit to the training data with a batch size of 100, 20% validation split, and 100 epochs

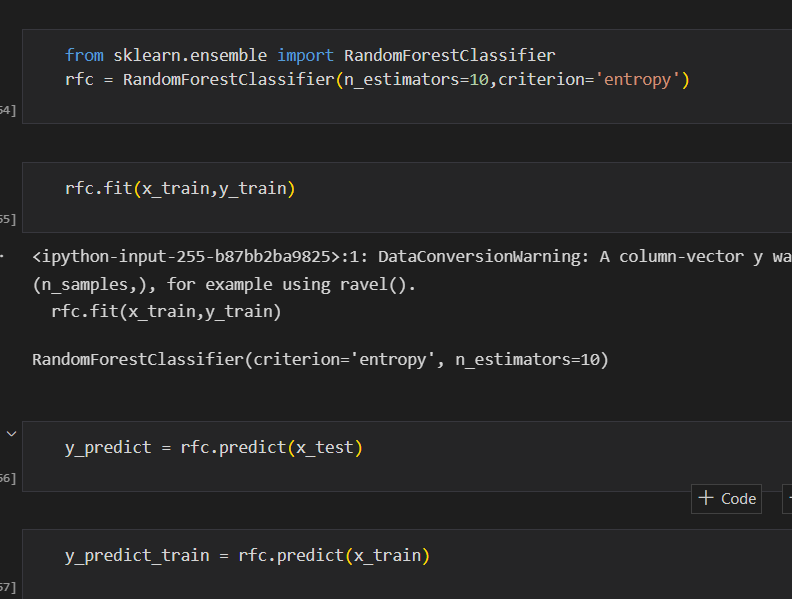






**Random Forest Model**

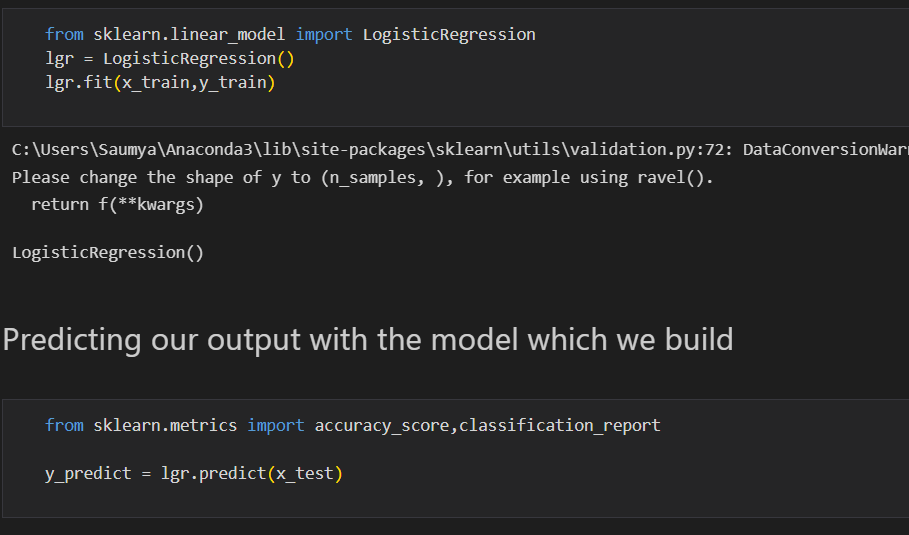
A function named random Forest is created and train and test data are passed as the parameters. Inside the function, Random Forest Classifier algorithm is initialised and training data is passed to the model with .fit() function. Test data is predicted with. predict() function and saved in a new variable. For evaluating the model, a confusion matrix and classification report is done.



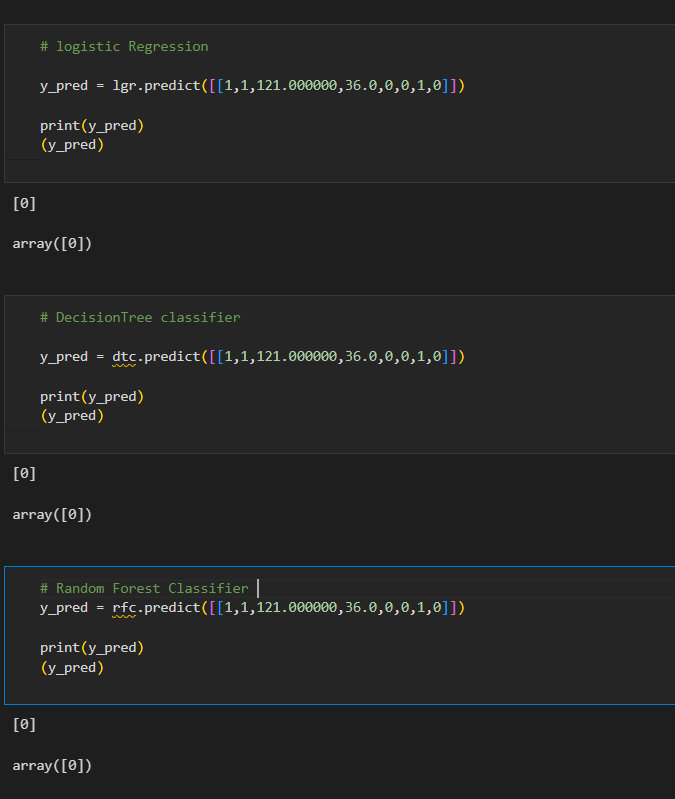
**Decision Tree Model**

A function named decision Tree is created and train and test data are passed as the parameters. Inside the function, Decision Tree Classifier algorithm is initialised and training data is passed to the model with the .fit() function. Test data is predicted with. predict() function and saved in a new variable. For evaluating the model, a confusion matrix and classification report is done.

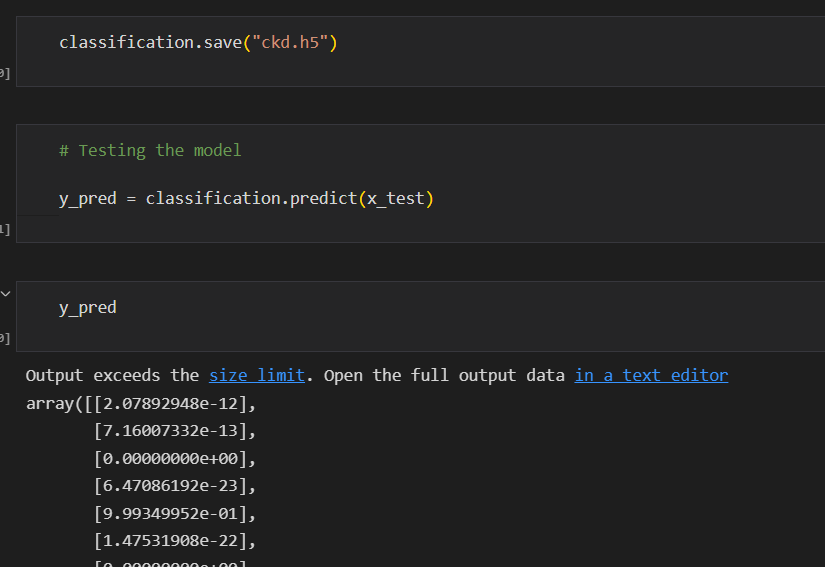
### Logistic Regression

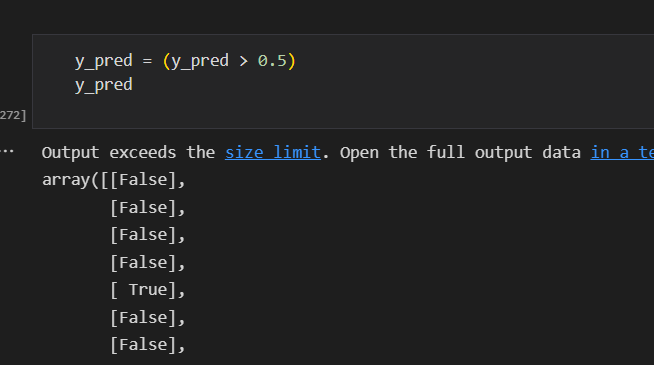
****

**Testing The Model**

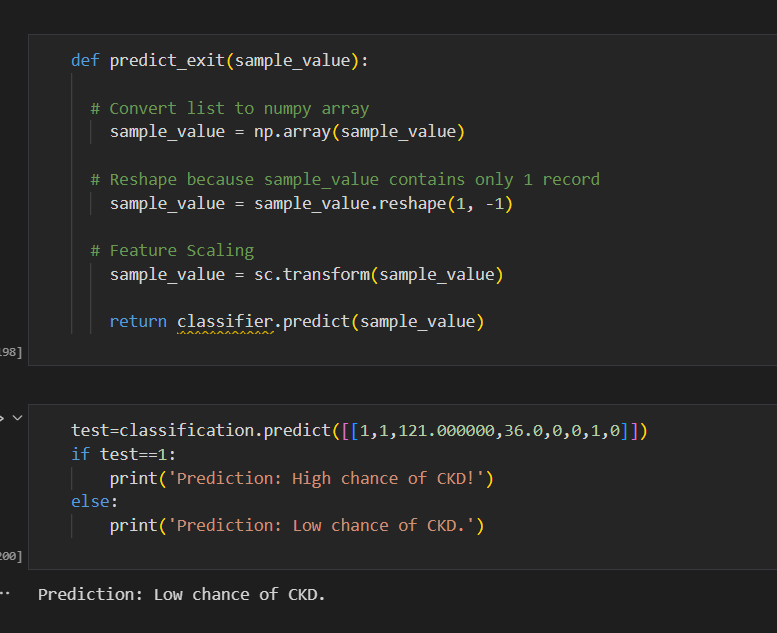
****

In ANN we first have to save the model to the test the inputs

****

****

This code defines a function named "predict\_exit" which takes in a sample\_value as an input. The function then converts the input sample\_value from a list to a numpy array. It reshapes the sample\_value array as it contains only one record. Then, it applies feature scaling to the reshaped sample\_value array using a scaler object 'sc' that should have been previously defined and fitted. Finally, the function returns the prediction of the classifier on the scaled sample\_value.

****

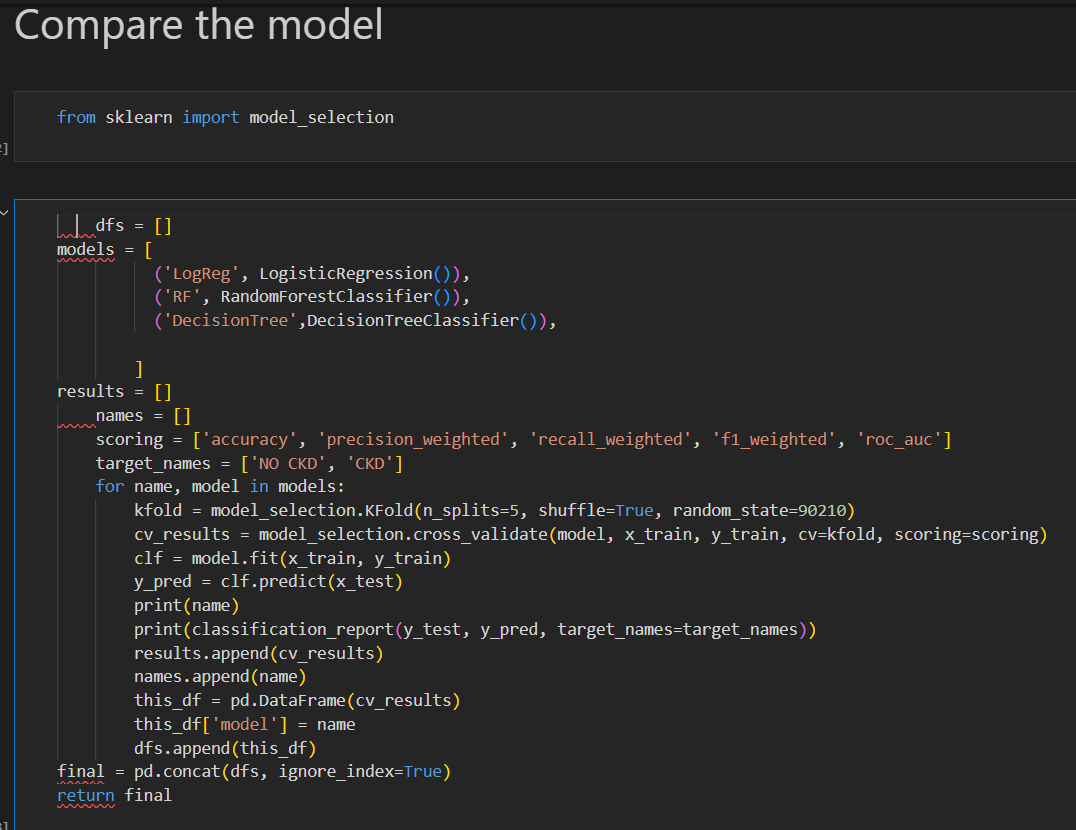
**6. CONCLUSION**

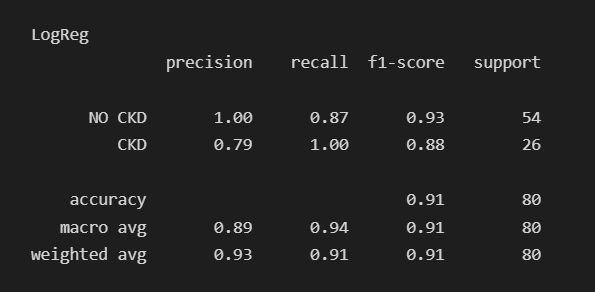
**Performance Testing & Evaluate The Results**

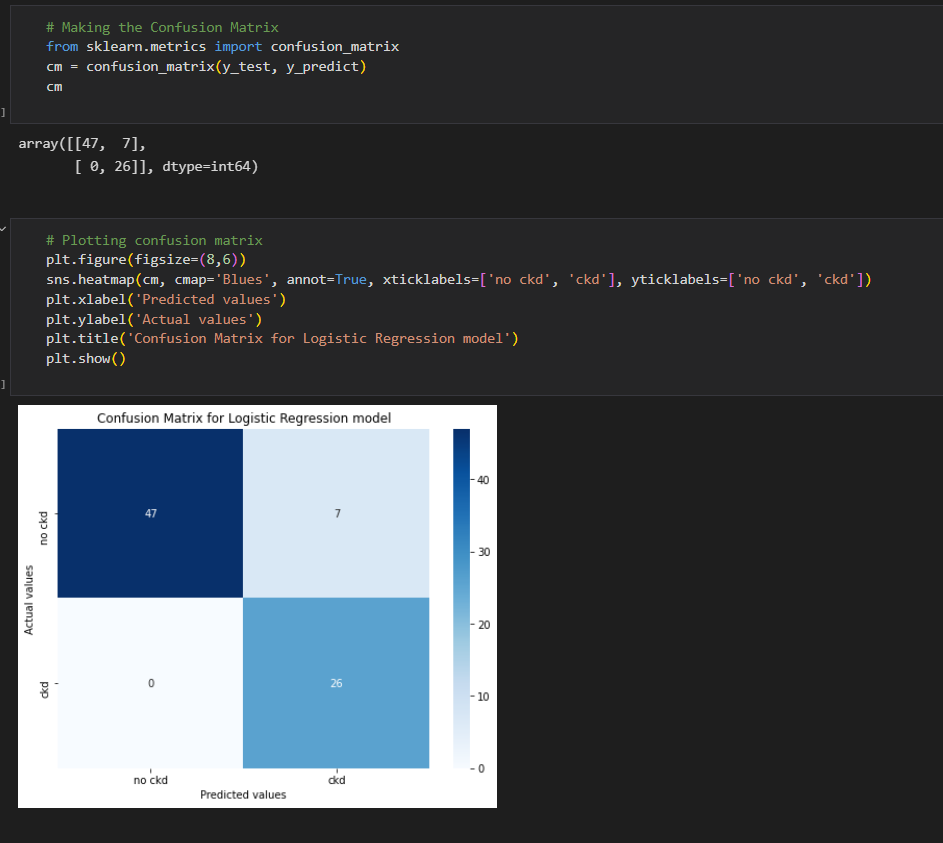
In this milestone, we will see the performance testing

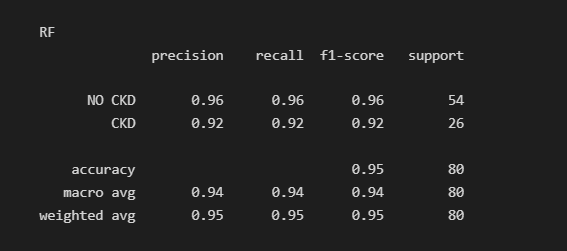
**Testing Model With Multiple Evaluation Metrics**

Multiple evaluation metrics means evaluating the model's performance on a test set using different performance measures. This can provide a more comprehensive understanding of the model's strengths and weaknesses. We are using evaluation metrics for classification tasks including accuracy, precision, recall, support and F1-score.

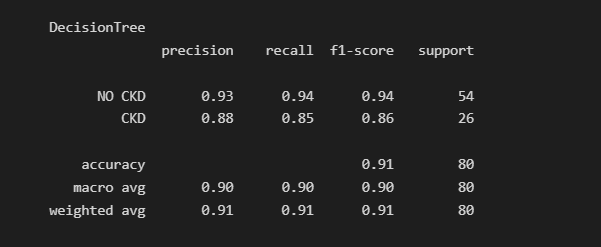


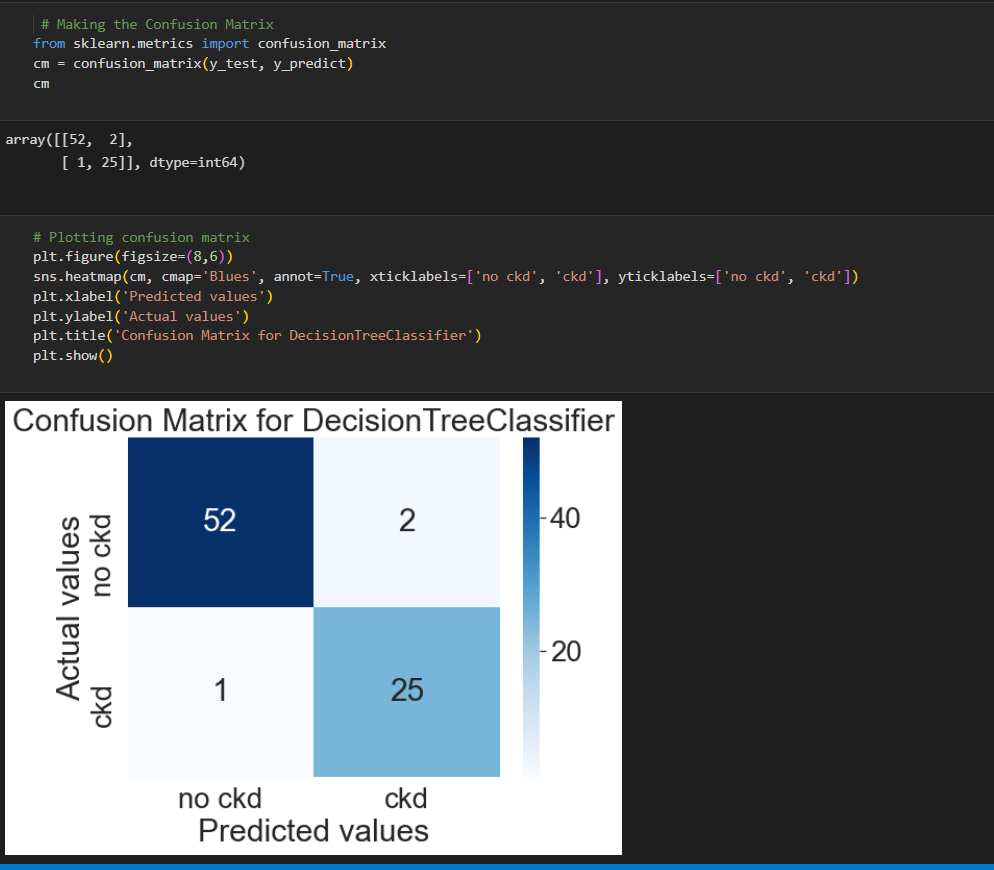
****

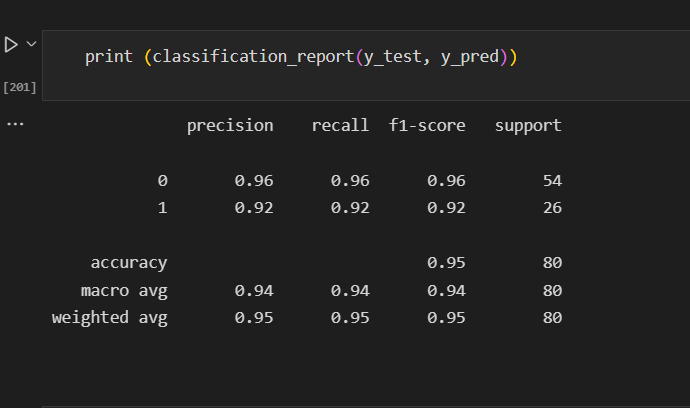
****

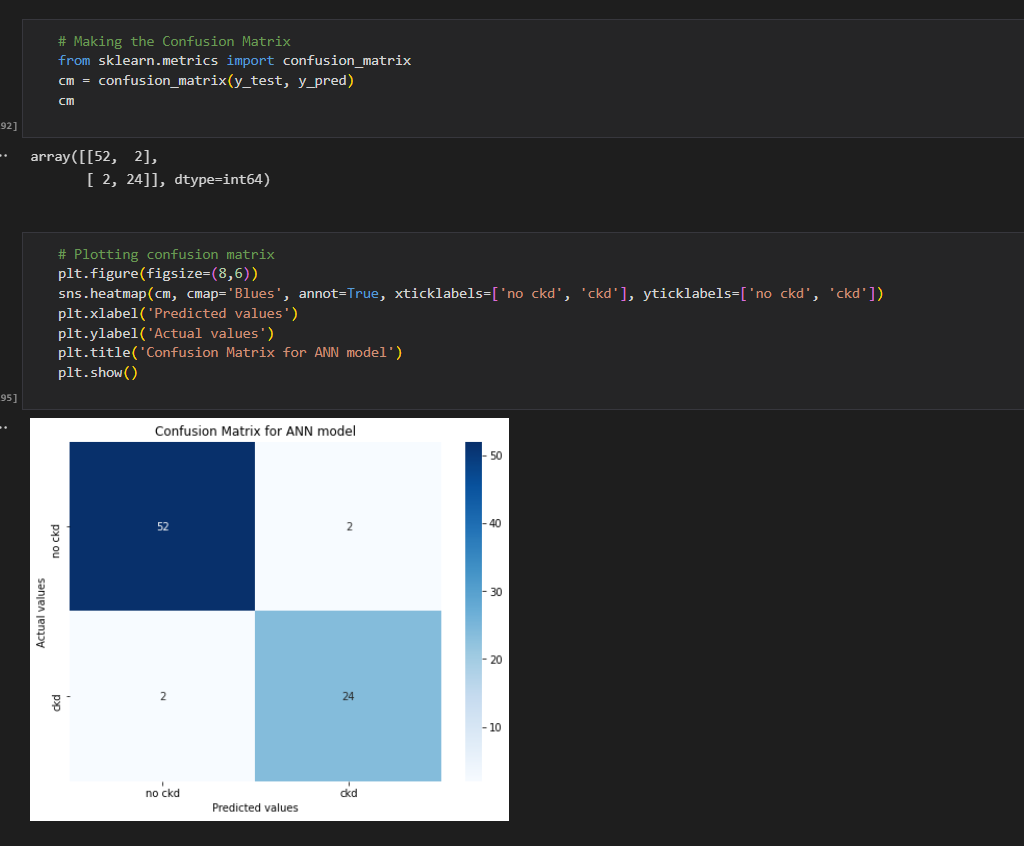
****

****

****

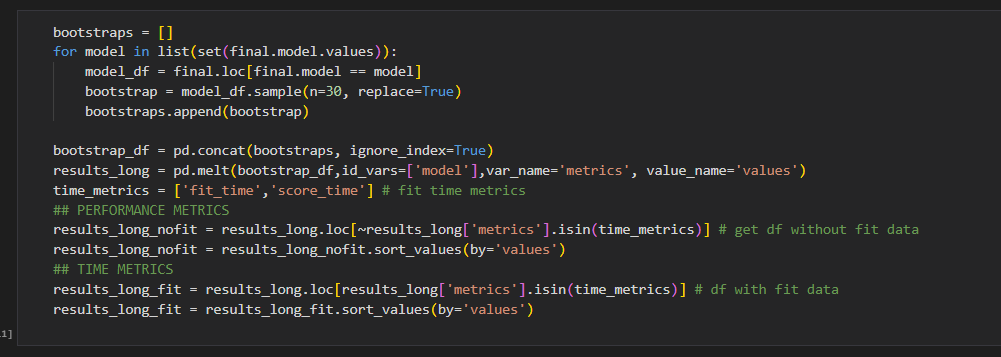
****

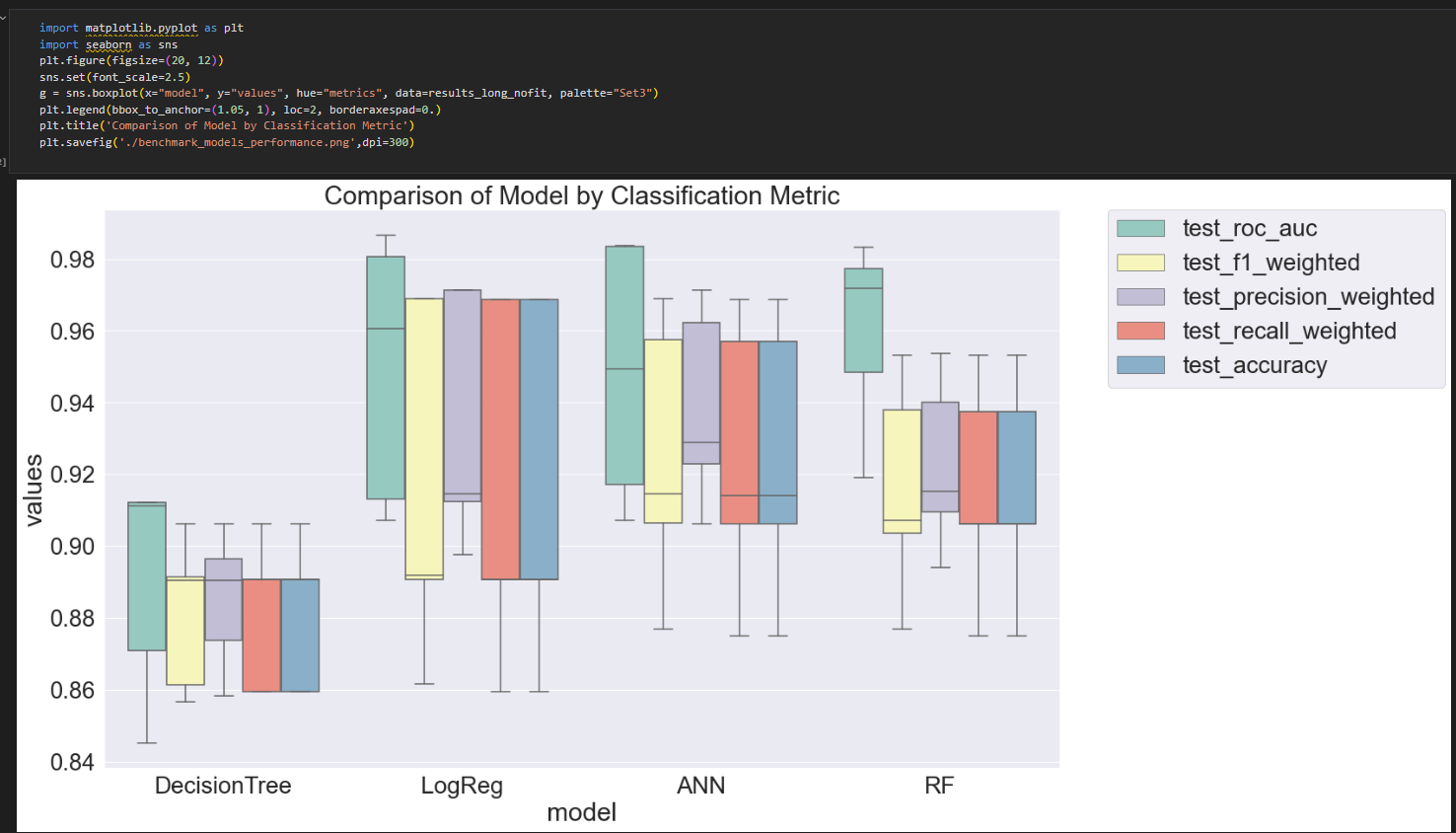
****

****

**A**ll above models are performing well for this dataset.

**Evaluate The Results**

****

****

**A**mong all these 4 models logistic regression has recall 1. So, we are going for logreg model.

**7. FUTURE SCOPE**

**Model Deployment**

In this milestone, we will see the model deployment.

**Save The Best Model**

Saving the best model after comparing its performance using different evaluation metrics means selecting the model with the highest performance and saving its weights and configuration. This can be useful in avoiding the need to retrain the model every time it is needed and to be able to use it in the future.



### Integrate With Web Framework

In this section, we will be building a web application that is integrated to the model we built. A UI is provided for the uses where he has to enter the values for predictions. The enter values are given to the saved model and prediction is showcased on the UI.

This section has the following tasks

* Building HTML Pages
* Building server-side script
* Run the web application

### Building Html Pages

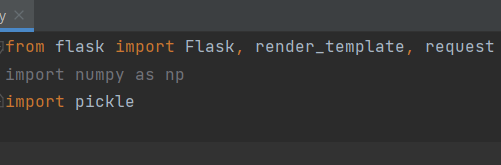
For this project create four HTML files namely

* home.html
* index1.html
* indexnew.html
* result.html

and save them in the templates folder

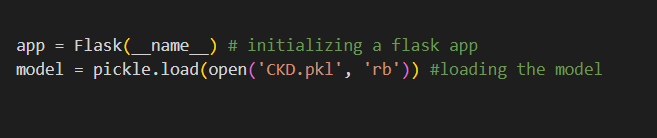
**Build Python Code**

Import the libraries

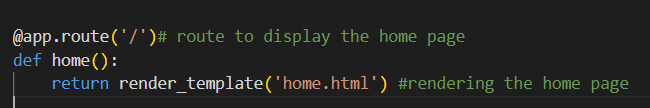
****

Load the saved model. Importing the flask module in the project is mandatory. An object of Flask class is our WSGI application.

Flask constructor takes the name of the current module (\_\_name\_\_) as argument.



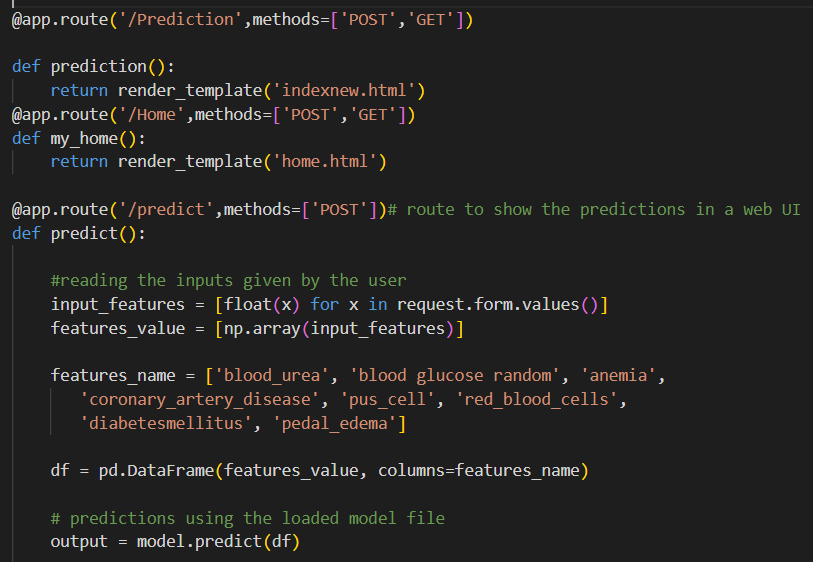
Render HTML page:



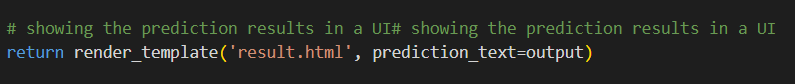
Here we will be using a declared constructor to route to the HTML page which we have created earlier.

In the above example, ‘/’ URL is bound with the home.html function. Hence, when the home page of the web server is opened in the browser, the html page will be rendered. Whenever you enter the values from the html page the values can be retrieved using POST Method.

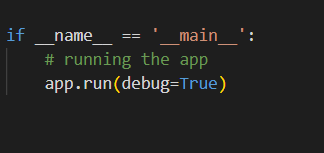
Retrieves the value from UI:

****

Here we are routing our app to predict() function. This function retrieves all the values from the HTML page using Post request. That is stored in an array. This array is passed to the model.predict() function. This function returns the prediction. And this prediction value will be rendered to the text that we have mentioned in the submit.html page earlier.

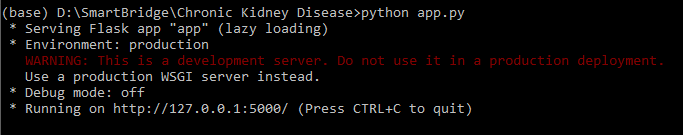
****

Main Function:

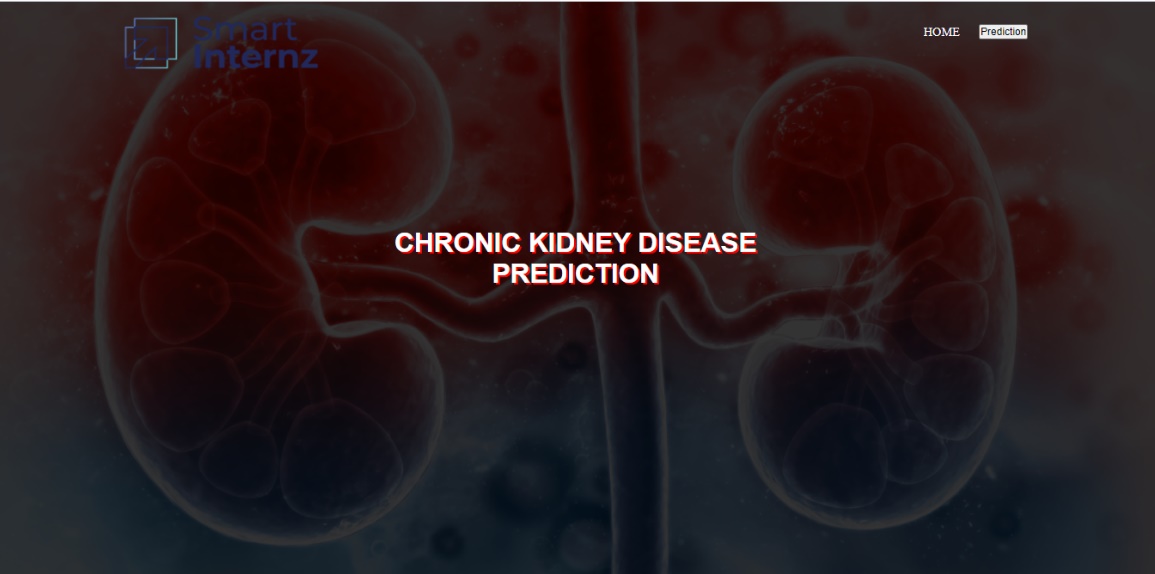


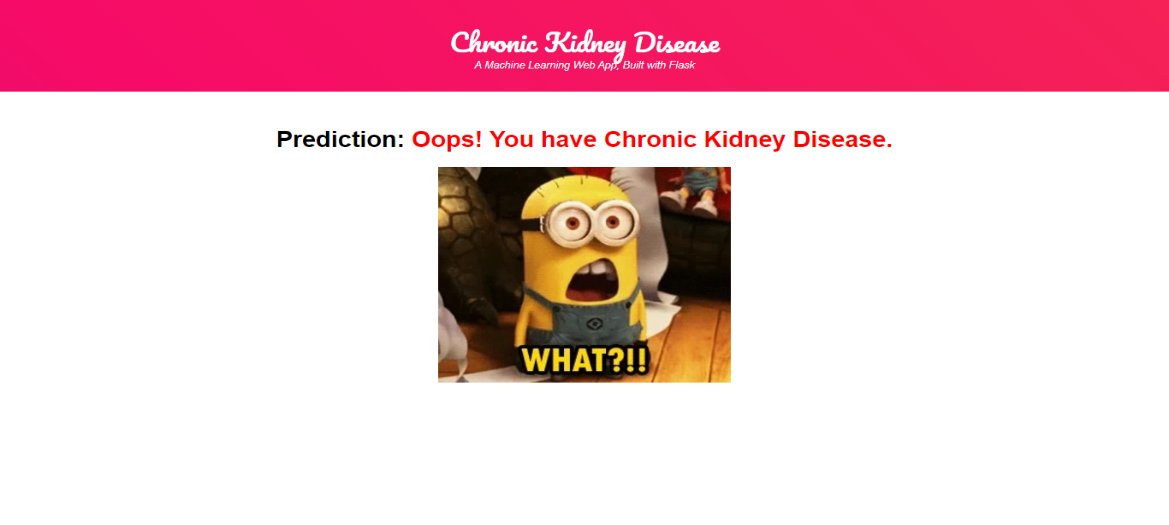
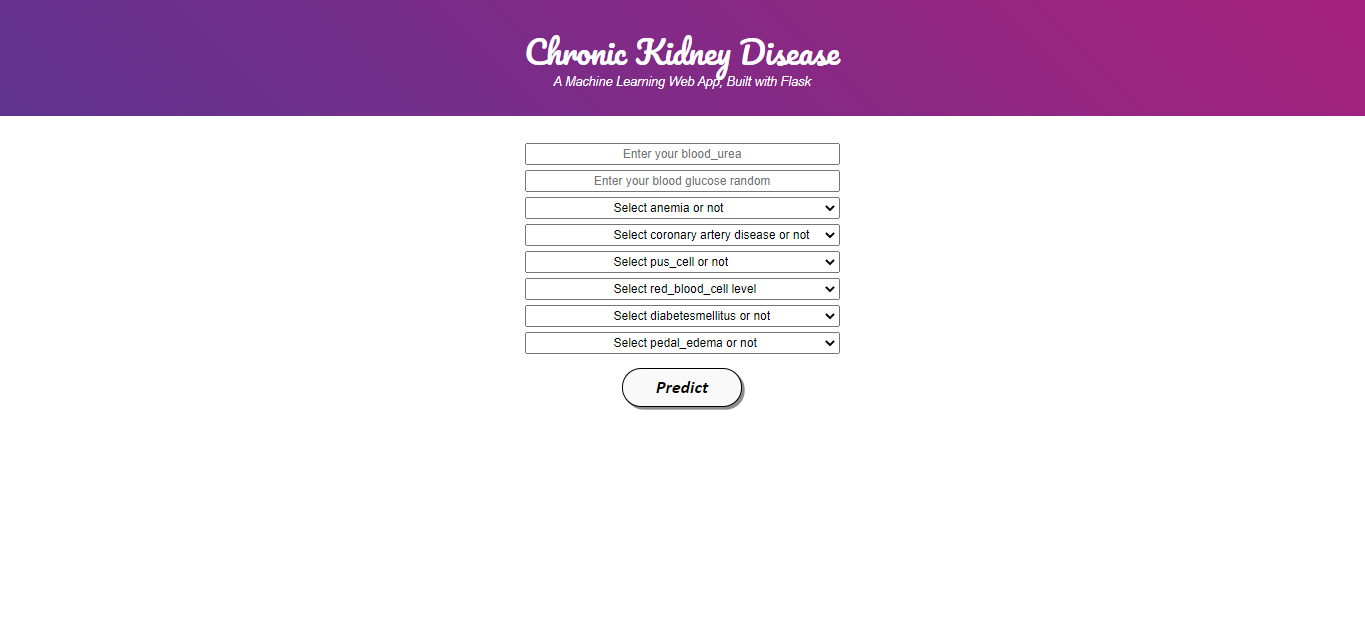
### Run The Web Application

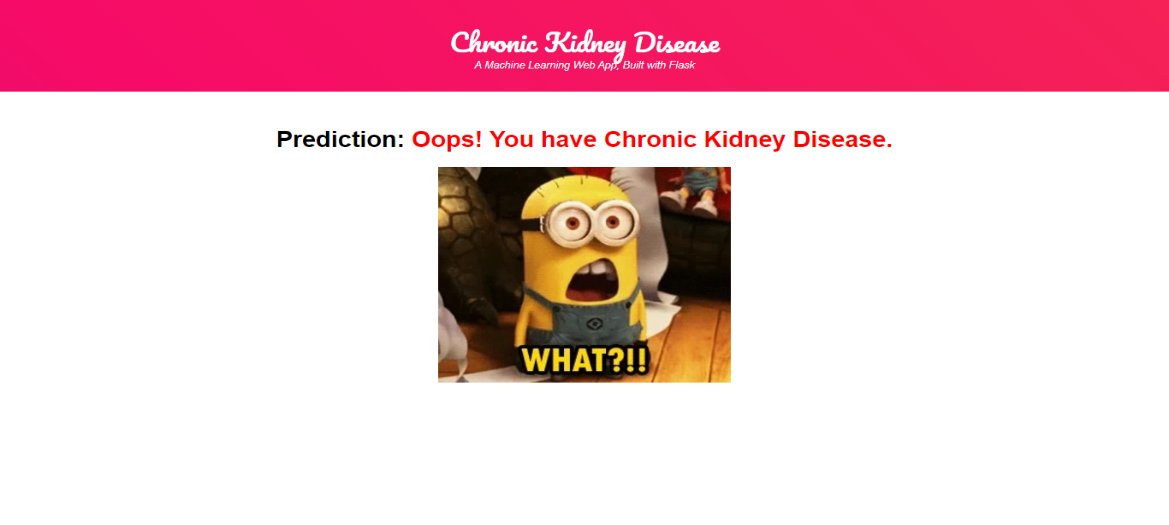
* Open anaconda prompt from the start menu
* Navigate to the folder where your python script is.
* Now type “python app.py” command
* Navigate to the localhost where you can view your web page.
* Click on the predict button from the top left corner, enter the inputs, click on the submit button, and see the result/prediction on the web.



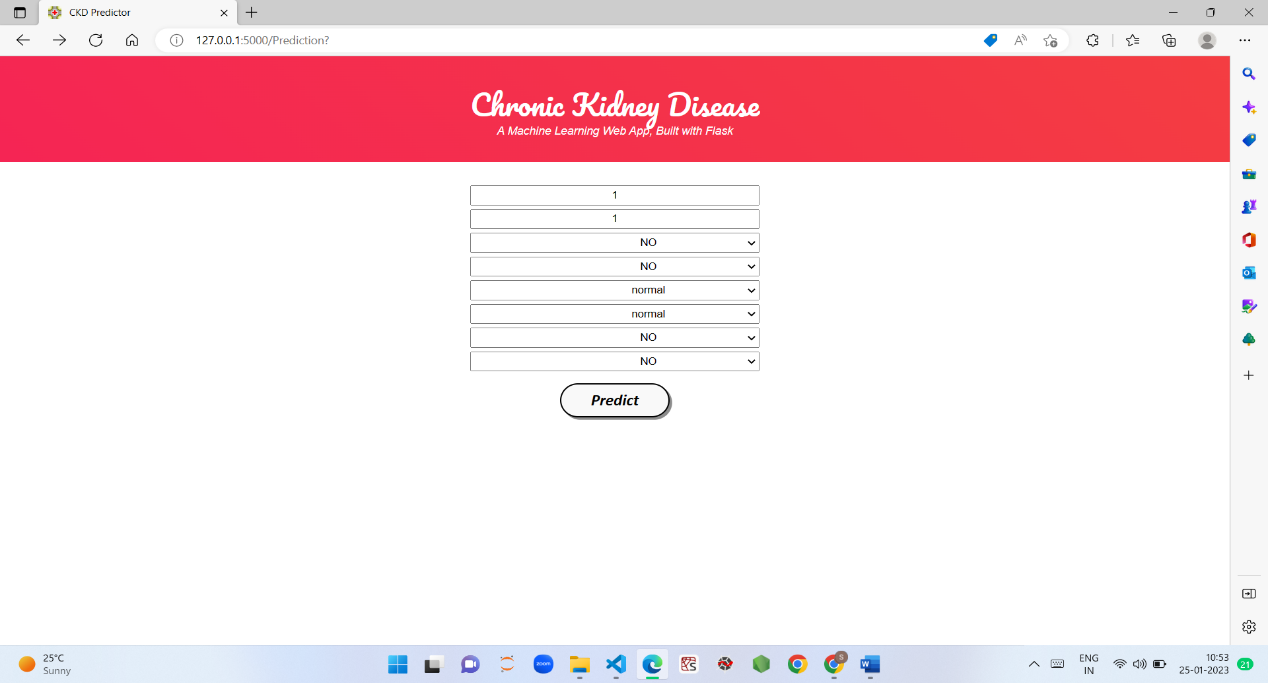
Now,Go the web browser and write the localhost url (http://127.0.0.1:5000) to get the below result

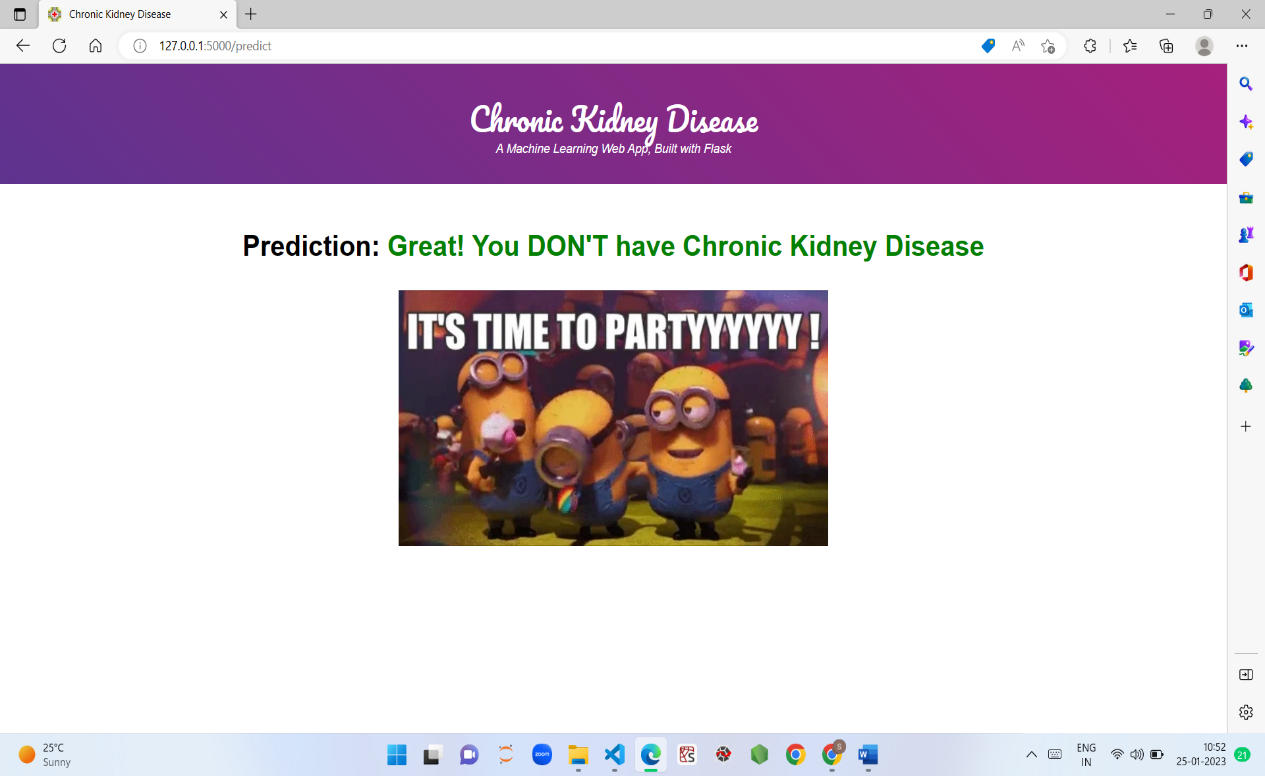






Input - Now, the user will give inputs to get the predicted result after clicking onto the submit button.

****

****

### Project Demonstration & Documentation

Project deliverables to be submitted along with other deliverables

**Record Explanation Video For Project End To End Solution**

Record the video explanation for end to end solution.

**Project Documentation-Step By Step Project Development Procedure**

Create document as per the template provided

**8.APPENDIX**

import numpy as np

import pandas as pd

from flask import Flask, request, render\_template

import pickle

app=Flask(\_\_name\_\_)

model=pickle.load(open('lgr.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=[float(x) for x in request.form.values()]

features\_value=[np.array(input\_features)]

features\_name=['blood\_urea','blood glucose random','coronary\_artery\_disease','anemia','pus\_cell','red\_blood\_cells','diabetesmellitus','pedal\_edema']

df=pd.DataFrame(features\_value, columns=features\_name)

output=model.predict(df)

render\_template('result.html',prediction\_text=output)

if \_\_name\_\_=='\_\_main\_\_':

app.run(debug=False)

