**Prediction of chronic kidney disease using Machine learning algorithm**

**INT 300 – INTERNSHIP-2 PROJECT REPORT**

***Submitted by***

**Marreddy Gayatri Devi – E0421053**

**Sivane Senthil – E0121027**

***In partial fulfilment for the award of the degree of***

**BACHELOR OF TECHNOLOGY**

**in**

**COMPUTER SCIENCE AND ENGINEERING**

**(Artificial Intelligence and Machine Learning)**

**Sri Ramachandra Faculty of Engineering and Technology**

**Sri Ramachandra Institute of Higher Education and Research, Porur, Chennai -600116**

**April, 2023**

**Prediction of chronic kidney disease using Machine learning algorithm**

**INT 300 – INTERNSHIP-2 PROJECT REPORT**

***Submitted by***

**Marreddy Gayatri Devi– E0421053**

**Sivane Senthil – E0121027**

***In partial fulfilment for the award of the degree of***

**BACHELOR OF TECHNOLOGY**

**in**

**COMPUTER SCIENCE AND ENGINEERING**

**(Artificial Intelligence and Machine Learning)**

**Sri Ramachandra Faculty of Engineering and Technology**

**Sri Ramachandra Institute of Higher Education and Research, Porur, Chennai -600116**

**April, 2023**

**BONAFIDE CERTIFICATE**

Certified that this project report **“Prediction of chronic kidney disease using machine learning algorithm”** is the bonafide record of work done by **“Marreddy Gayatri devi–E0421053”** who carried out the internship work under my supervision.

**Signature of the Supervisor Signature of Programme Coordinator**

|  |  |
| --- | --- |
| Dr. Shiyamala Gowri  Assistant Professor,  Department of Artificial Intelligence and Data Analytics  Sri Ramachandra Faculty of Engineering and Technology,  SRIHER, Porur, Chennai-600 116. | Namasivaya Naveen S  Lecturer,  Department of Computer science and Medical Engineer  Sri Ramachandra Faculty of Engineering and Technology,  SRIHER, Porur, Chennai-600 116. |

**Evaluation Date:**

**BONAFIDE CERTIFICATE**

Certified that this project report **“Prediction of chronic kidney disease using machine learning algorithm”** is the bonafide record of work done by **“Sivane Senthil–E0121027 “**who carried out the internship work under my supervision.

**Signature of the Supervisor Signature of Programme Coordinator**

|  |  |
| --- | --- |
| Dr. Shiyamala Gowri  Assistant Professor,  Department of Artificial Intelligence and Data Analytics  Sri Ramachandra Faculty of Engineering and Technology,  SRIHER, Porur, Chennai-600 116. | Dr. Vanitha Magesh  Assistant Professor,  Department of Computer science and Machine learning  Sri Ramachandra Faculty of Engineering and Technology,  SRIHER, Porur, Chennai-600 116. |

**Evaluation Date:**

**ACKNOWLEDGEMENT**

I express my sincere gratitude to our Programme Coordinator **Prof.** **Namasivaya Naveen S** for their support and for providing the required facilities for carrying out this study.

I wish to thank my faculty supervisor(s), **prof. shiyamala Gowri** Department of Artificial Intelligence and Data analytics, Sri Ramachandra faculty of Engineering and Technology and for extending help and encouragement throughout the project. Without his/her continuous guidance and persistent help, this project would not have been a success for me.

I am grateful to all the members of Sri Ramachandra Faculty of Engineering and Technology, my beloved parents and friends for extending the support, who helped us to overcome obstacles in the study.

**ACKNOWLEDGEMENT**

I express my sincere gratitude to our Programme Coordinator **Dr. Vanitha Magesh** for their support and for providing the required facilities for carrying out this study.

I wish to thank my faculty supervisor(s), **prof. shiyamala Gowri** Department of Artificial Intelligence and Data analytics, Sri Ramachandra faculty of Engineering and Technology and for extending help and encouragement throughout the project. Without his/her continuous guidance and persistent help, this project would not have been a success for me.

I am grateful to all the members of Sri Ramachandra Faculty of Engineering and Technology, my beloved parents, and friends for extending the support, who helped us to overcome obstacles in the study.

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **CHAPTER NO.** | **TITLE** | **PAGE NO.** |
|  | **ABSTRACT** | **8** |
|  |  |  |
| **1** | **INTRODUCTION**  **1.1 PROBLEM STATMENT** | **9** |
| **2** | **LITERATURE REVIEW** | **11** |
| **3** | **METHODOLOGY** | **12** |
| **4** | **IMPLEMENTATION**  **4.1 MACHINE LEARNING**  **4.2 ALGORITHMS** | **13** |
| **5**  **6**  **7**  **8**  **9** | **APPENDICES**  **5.1 APPENDIX-1: CODE COMPILER**  **5.2 APPENDIX-2: SCREENSHOT**  **RESULTS & DISCUSSIONS**  **CONCLUSION & FUTURE**  **REFERENCES**  **WORKLOG** | **19**    **34**  **35**  **36**  **38** |
|  |  |  |

**ABSTRACT**

The purpose of this study was to assess the feasibility of machine learning in predicting the risk of end-stage kidney disease from patients with chronic kidney disease. Chronic kidney disease is a life-threatening condition that can be difficult to diagnose early because there are no symptoms. The purpose of the proposed study is to develop and validate a predictive model for the prediction of chronic kidney disease. Machine learning algorithms are often used in medicine to predict and classify diseases. Medical records are often skewed. We have used chronic kidney disease dataset from Data world with 25 features and 400 attributes.

**CHAPTER 1**

**INTRODUCTION**

Chronic kidney disease (CKD) is a significant healthcare burden that affects billions of individuals worldwide and makes a profound impact on global mortality. In the United States, approximately 11% of the population or 37 million people suffer from CKD that results in an annual Medicare cost of $84 billion. The prevalence of this disease is estimated at 10.8% in China, affecting about 119.5 million people.

Gradual loss of the kidney function can lead to end stage kidney disease in CKD patients, precipitating the need for kidney replacement therapy. Timely intervention in those CKD patients who have a high risk of ESKD may not only improve these patient’s quality of life by delaying the disease progression, but also reduce the mortality and healthcare. Because the disease progression is typically silent, a reliable prediction model for risk of ESKD at the early stage of CKD can be clinically essential.

A few statistical models were developed to predict the likelihood of ESKD based on certain variables, including age, lab results, and most commonly, the estimated albumin.

With the advent of the big data era, new methods became available in developing a predictive model that used to rely on traditional statistics. Machine learning (ML) is a subset of artificial intelligence (AI) that allows the computer to perform a specific task without explicit instructions. When used in predictive modelling, ML algorithm can be trained to capture the underlying patterns of the sample data and make predictions about the new data based on the acquired information

**1.1 PROBLEM STATEMENT**

The major challenge in chronic kidney disease is its detection. There are instruments available which can predict chronic kidney disease but either they are expensive or are not efficient to calculate the chance of chronic kidney disease in humans. Early detection of kidney diseases can decrease the mortality rate and overall complications. However, it is not possible to monitor patients every day in all cases accurately and consultation of a patient for 24 hours by a doctor is not available since it requires more patience, time, and expertise. Since we have a good amount of data in today’s world, we can use various machine learning algorithms to analyse the data for hidden patterns. The hidden patterns can be used for health diagnosis in medicinal data**.**

**CHAPTER 2**

**LITRETURE SURVEY**

The reviews indicate that several studies have been conducted on chronic kidney disease prediction using machine-learning techniques. There are various parameters which play an important role in improving model performance like dataset size, quality of dataset and the time dataset collected.

The focus on chronic kidney disease prediction using machine learning models based on the dataset with a larger size aims to predict Three machine learning algorithms;

Random Forest, Support Vector Machine and Decision Tree.

The algorithms were selected based on their popularity in chronic kidney disease prediction and their performance of classification on previous research works

Such as Priyanka etcarried out chronic kidney disease prediction through naive bayes. They have tested using other algorithms such as KNN (K-Nearest Neighbour Algorithm), SVM (Support Vector Machines), Decision tree, and ANN (Artificial Neural Network) and they have got Naïve Bayes with better accuracy of 94.6% when compared to other algorithms.

Similarly, Random Forest has the highest voting results of the target output

**CHAPTER 3**

**METHODOLOGY**

Once the data is collected, it then enters the data preparation stage. Data preparation, often referred to as “pre-processing” is the stage at which raw data is cleaned up and organized for the following stage of data processing.

Data Pre-processing in machine learning is a crucial step that helps enhance the quality of data to promote the extraction of meaningful insights from the data. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviours or trends, and is likely to contain many errors. Data pre-processing is a proven method of resolving such issues. It converts raw data into understandable and readable data. Data Pre-processing involves three stages - Data Cleaning, Data pre-processing, and Data Reduction.

* Data cleaning is an essential step in preparing data for machine learning. Clears data through methods such as Fixing Missing Values and Filling Missing Value.
* Data pre-processing in Machine Learning refers to the technique of preparing (cleaning and organizing) the raw data to make it suitable for a building and training Machine Learning models.
* Data reduction when managing large databases, analysis is too difficult. As a result, we eliminate any independent variables that have little, or they will have no effect the target variable

**CHAPTER 4**

**IMPLEMENTATION**

● Train Dataset: Used to fit the machine learning model.

● Test Dataset: Used to evaluate the fit machine learning model.

test size:

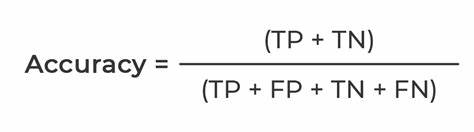
If float, should be between 0.0 and 1.0 and represent the proportion of the dataset to include in the test split. If int, represents the absolute number of test samples. If None, the value is set to the complement of the train size. If train size is also None, it will be set to 0.25.

train size:

If float, should be between 0.0 and 1.0 and represent the proportion of the dataset to include in the train split. If int, represents the absolute number of train samples. If None, the value is automatically set to the complement of the test size.

Accuracy

Accuracy implies the ability of the classification algorithm to predict the classes of the dataset correctly. It is the measure of how close or near the predicted value is to the actual or theoretical value. Generally, accuracy is the measure of the ratio of correct predictions over the total number of instances.



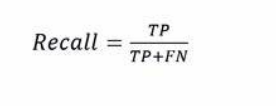
Precision

Precision measures the true values correctly predicted from the total predicted values in the actual class. Precision quantifies the ability of the classifiers to not label a negative example as positive.



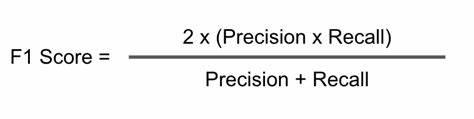
Recall

Recall measures the rate of positive values that are correctly classified. Recall answers the question of what proportion of actual positives are correctly classified.



F-measure

F-measure is also called F1-score is the harmonic mean between recall and precision.



**4.1 MACHINE LEARNING**

In machine learning, classification refers to a predictive modelling problem where a class label is predicted for a given example of input data.

● Supervised Learning

Supervised learning is the type of machine learning in which machines are trained using well "labelled" training data, and on the basis of that data, machines predict the output. The labelled data means some input data is already tagged with the correct output. In supervised learning, the training data provided to the machines work as the supervisor that teaches the machines to predict the output correctly. It applies the same concept as a student learns in the supervision of the teacher. Supervised learning is a process of providing input data as well as correct output data to the machine learning model. The aim of a supervised learning algorithm is to find a mapping function to map the input variable(x) with the output variable(y).

● Unsupervised learning

Unsupervised learning cannot be directly applied to a regression or classification problem because unlike supervised learning, we have the input data but no corresponding output data. The goal of unsupervised learning is to find the underlying structure of the dataset, group that data according to similarities, and represent that dataset in a compressed format.

• Unsupervised learning is helpful for finding useful insights from the data.

• Unsupervised learning is much similar to how a human learns to think by their own experiences, which makes it closer to the real AI.

• Unsupervised learning works on unlabelled and uncategorized data which make unsupervised learning more important.

• In real-world, we do not always have input data with the corresponding output so to solve such cases, we need unsupervised learning.

● Reinforcement learning

Reinforcement learning is an area of Machine Learning. It is about taking suitable action to maximize reward in a particular situation. It is employed by various software and machines to find the best possible behaviour or path it should take in a specific situation. Reinforcement learning differs from supervised learning in a way that in supervised learning the training data has the answer key with it so the model is 11 trained with the correct answer itself whereas in reinforcement learning, there is no answer but the reinforcement agent decides what to do to perform the given task. In the absence of a training dataset, it is bound to learn from its experience.

**4.2 ALGORITHMS**

Linear regression: Linear regression analysis is used to predict the value of a variable based on the value of another variable.

Polynomial regression: Polynomial Regression is a regression algorithm that models the relationship between a dependent(y) and independent variable(x) as nth degree polynomial.

Decision tree regression: Decision tree regression observes features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output.

Random forest regression: It builds decision trees on different samples and takes their average in case of regression

SVR: Support Vector Regression is a supervised learning algorithm that is used to predict discrete values. Support Vector Regression uses the same principle as the SVMs. The basic idea behind SVR is to find the best fit line. In SVR, the best fit line is the hyperplane that has the maximum number of points.

Naive Bayes: It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.

P(A|B) = P(B|A) \* P(A) / P(B)

KNN: K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.

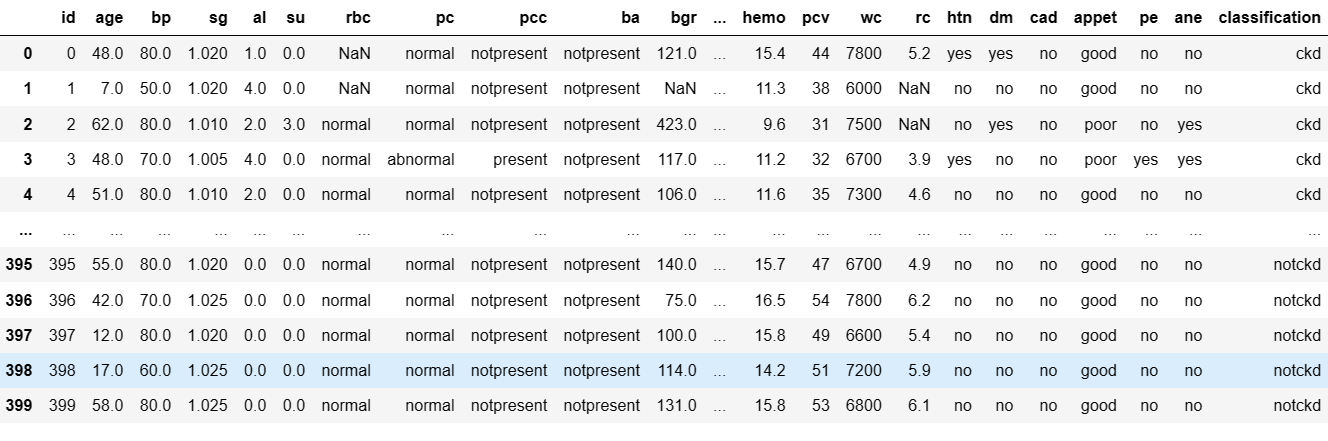
Decision tree classification: It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.

It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome

Random forest classification: It builds decision trees on different samples and takes their majority vote for classification.

SVM: The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called support vectors, and hence the algorithm is termed as Support Vector Machine.

**DATASET**

****

**APPENDICES**

**APPENDIX-1: CODE COMPILER**

**CLASSIFICATIONS**

# **LOGISTIC REGRESSION**

from sklearn.linear\_model import LogisticRegression

from sklearn.metrics import accuracy\_score

from sklearn.model\_selection import train\_test\_split

import pandas as pd

*# Load data into a Pandas DataFrame*

data = pd.read\_csv('kidney\_disease.csv')

*# Separate the input (X) and output (y) variables*

X = df.iloc[:, :-1].values

y = df.iloc[:, -1].values

*# Split the data into training and testing sets*

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=0)

*# Create a Logistic Regression object and fit it to the training data*

Classifier = LogisticRegression()

Classifier.fit(X\_train, y\_train)

*# Predict the output values for the testing set*

y\_pred = Classifier.predict(X\_test)

*# Calculate the accuracy score*

accuracy = accuracy\_score(y\_test, y\_pred)

​*# Print the accuracy score*

print("Accuracy:", accuracy)

**REGRESSION**

# **RANDOM FOREST REGRESSION**

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn.model\_selection import train\_test\_split

from sklearn.ensemble import RandomForestClassifier

from sklearn.metrics import accuracy\_score, confusion\_matrix, classification\_report

X = df.iloc[:, :-1].values

y = df.iloc[:, -1].values

#Instantiate and fit the RandomForestClassifier

forest = RandomForestClassifier()

forest.fit(X\_train, y\_train)

# Make predictions for the test set

y\_pred\_test = forest.predict(X\_test)

# View accuracy score

accuracy\_score(y\_test, y\_pred\_test)

# View confusion matrix for test data and predictions

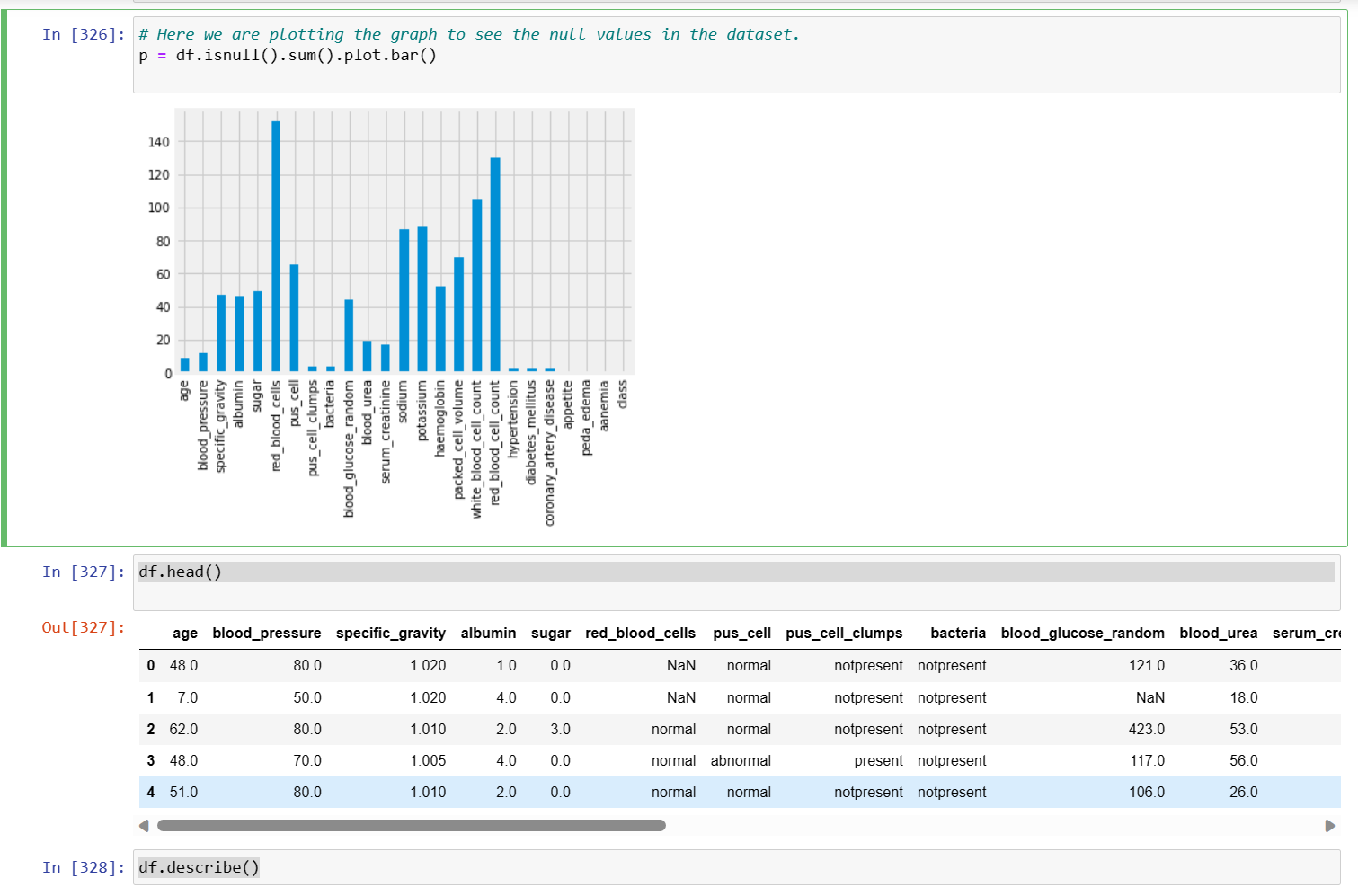
confusion\_matrix(y\_test, y\_pred\_test)

# View the classification report for test data and predictions

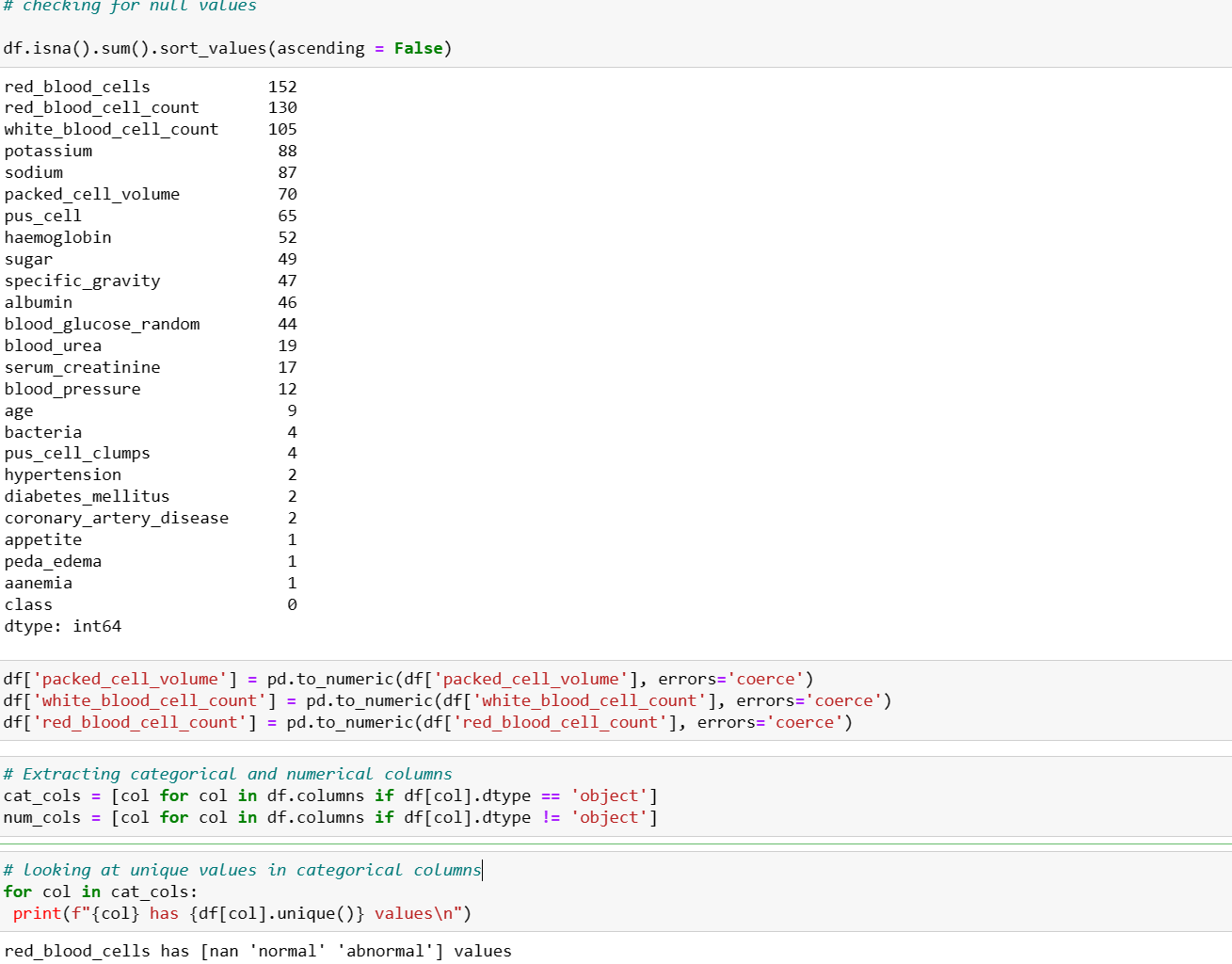
print(classification\_report(y\_test, y\_pred\_test)

**APPENDIX- 2**

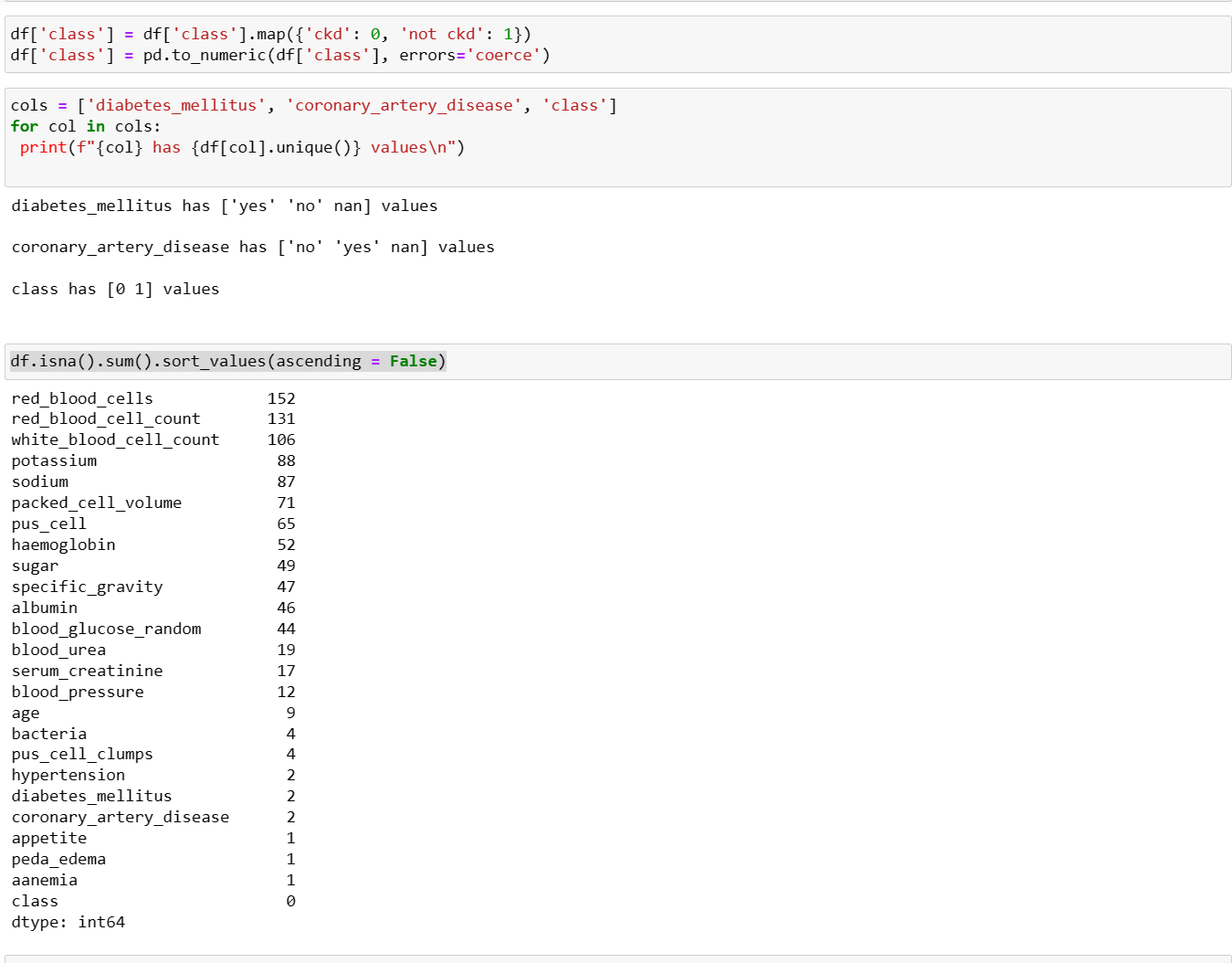
**SCREENSHOTS**

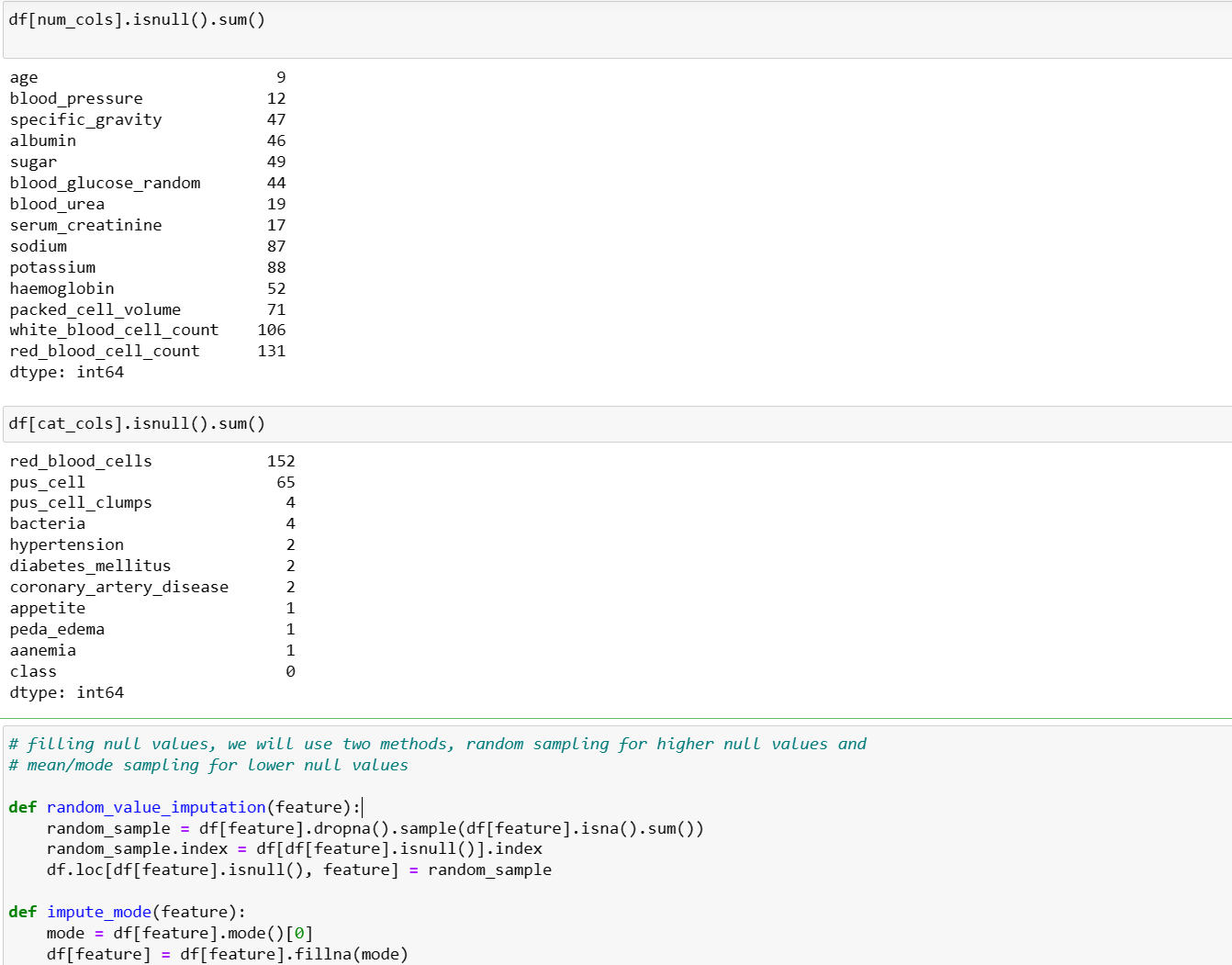


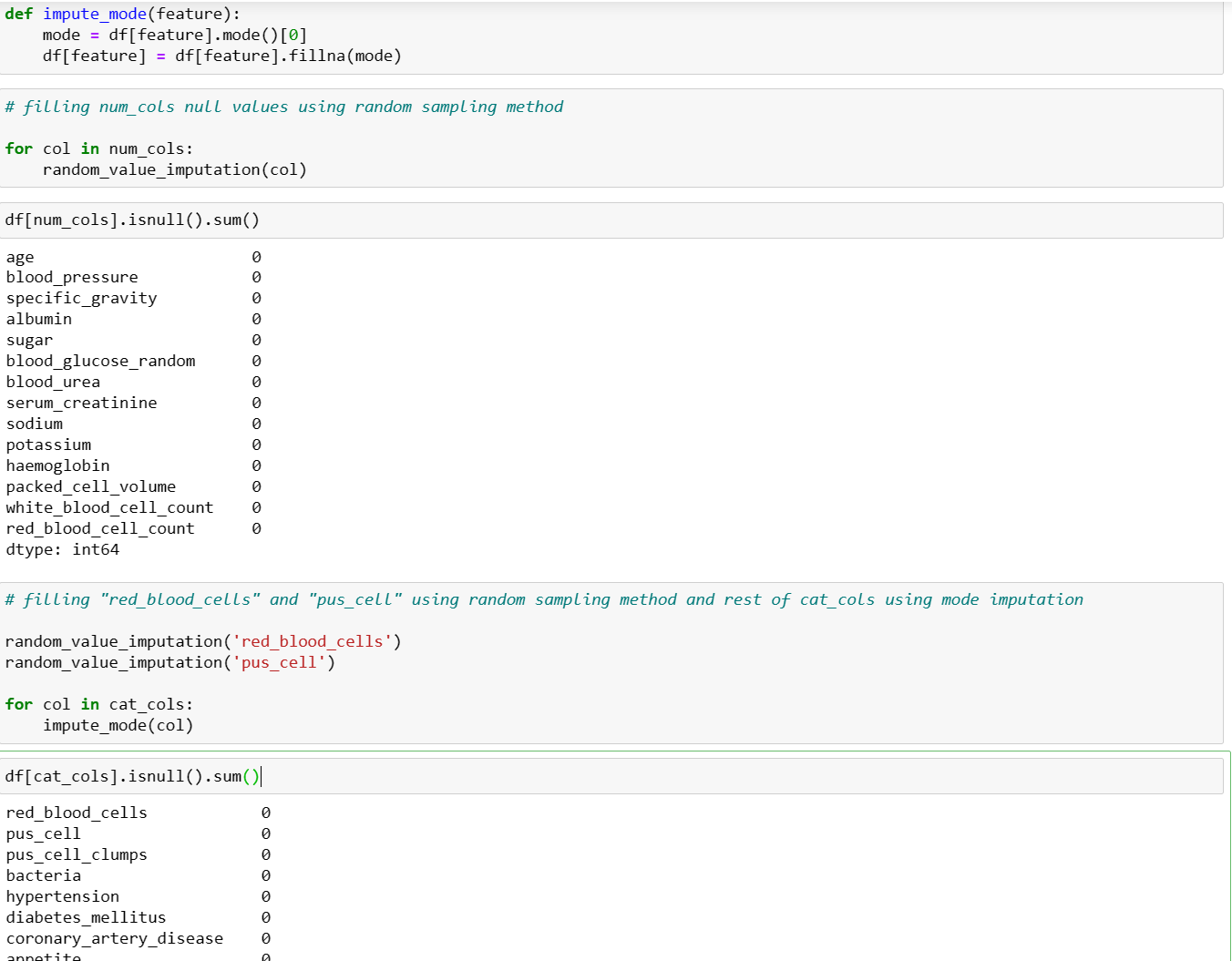
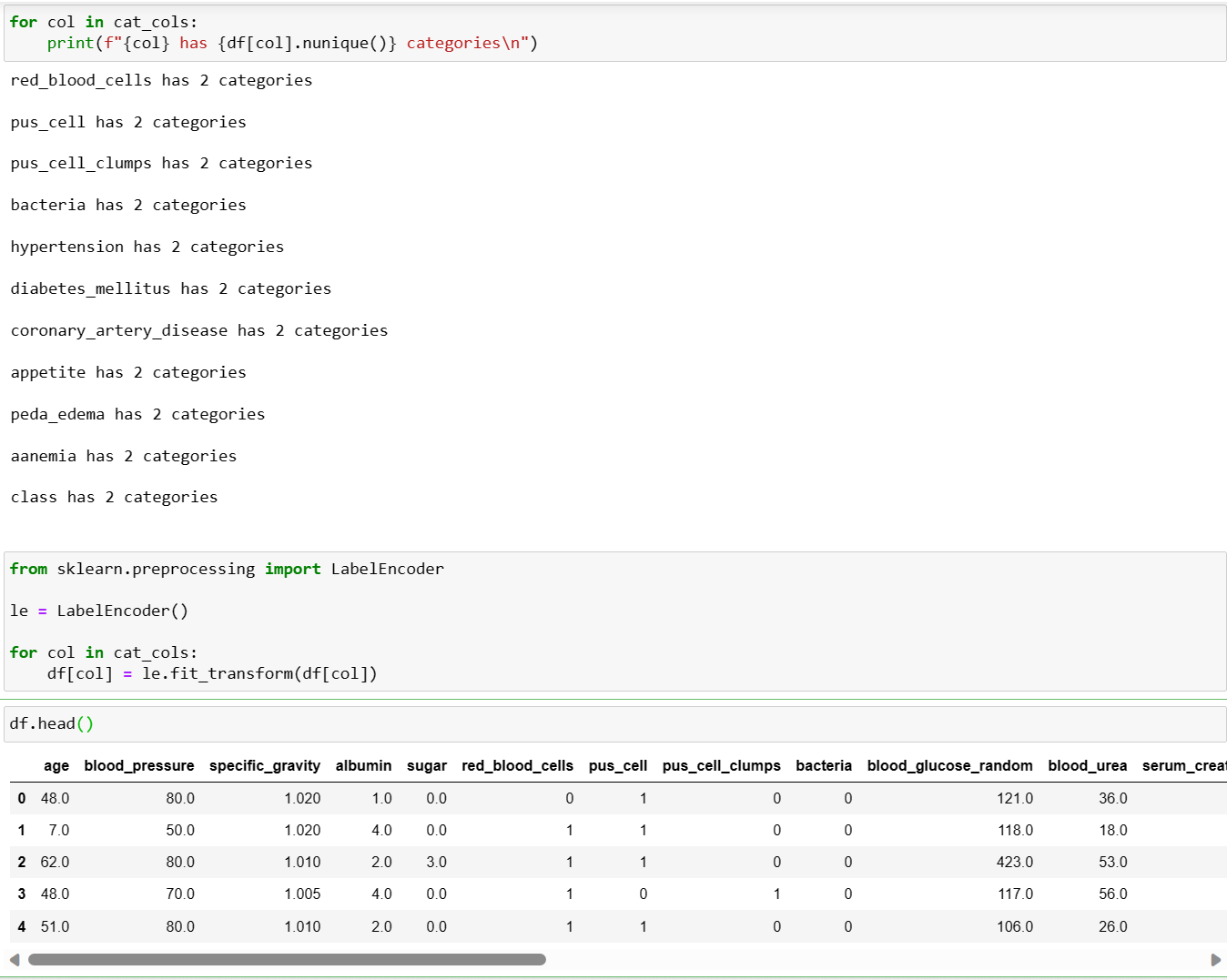
****

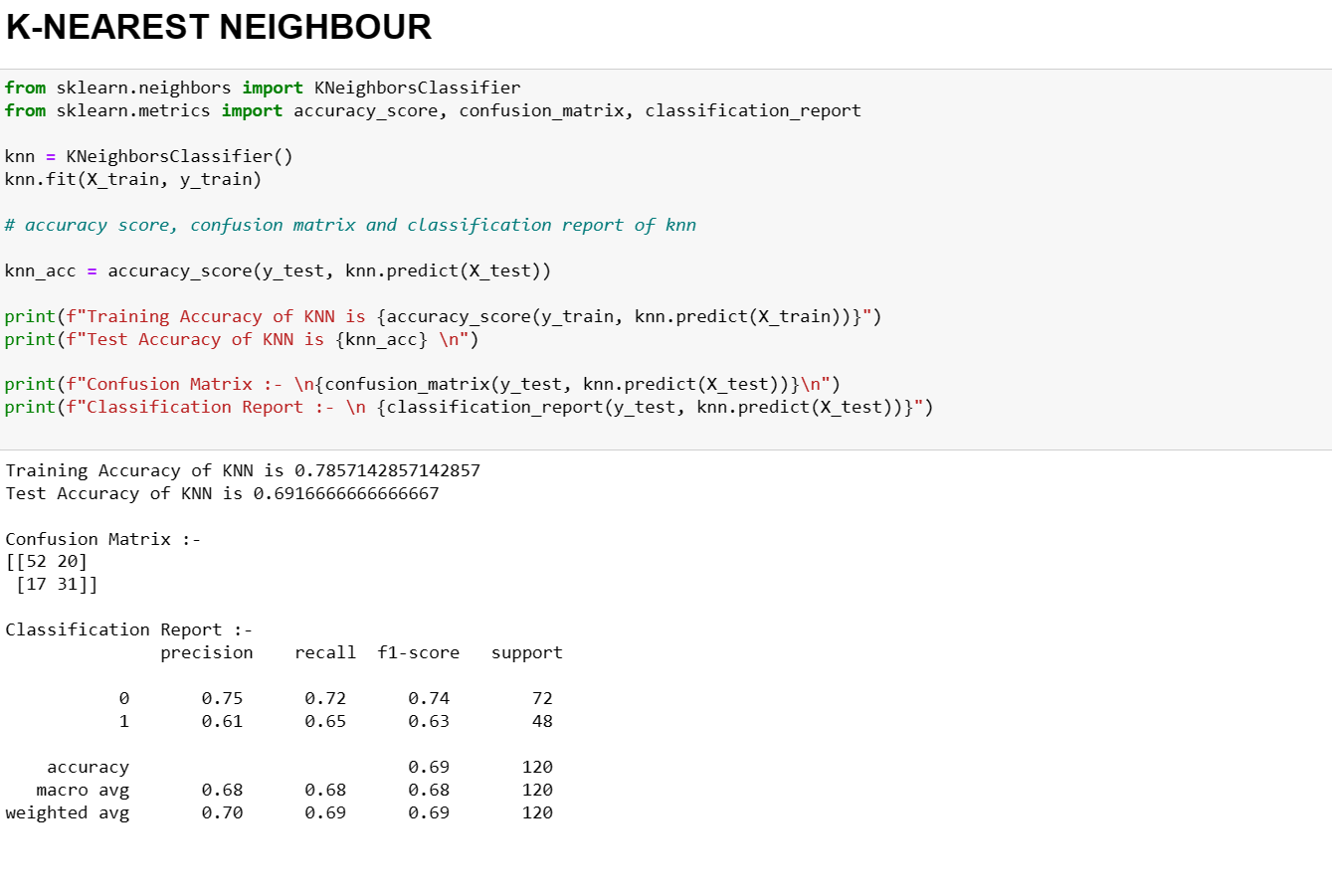


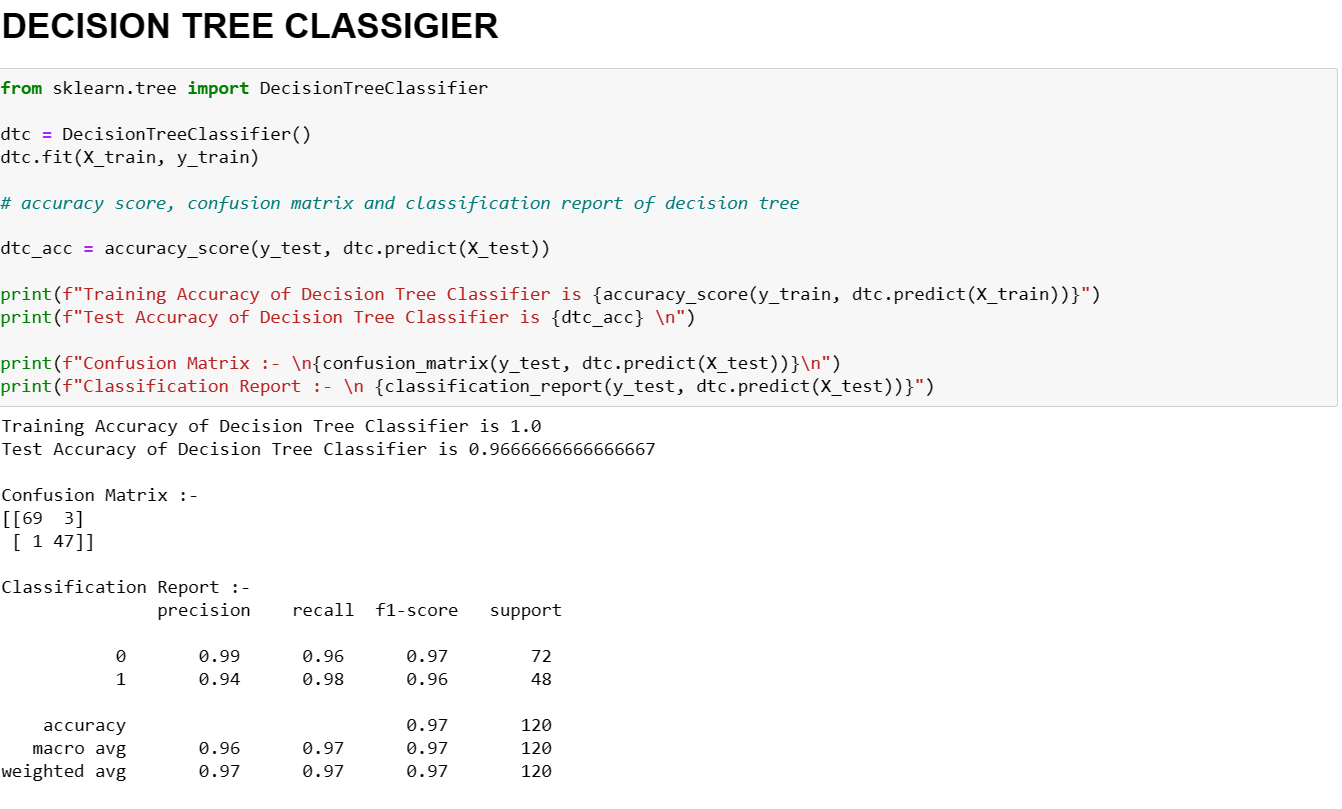


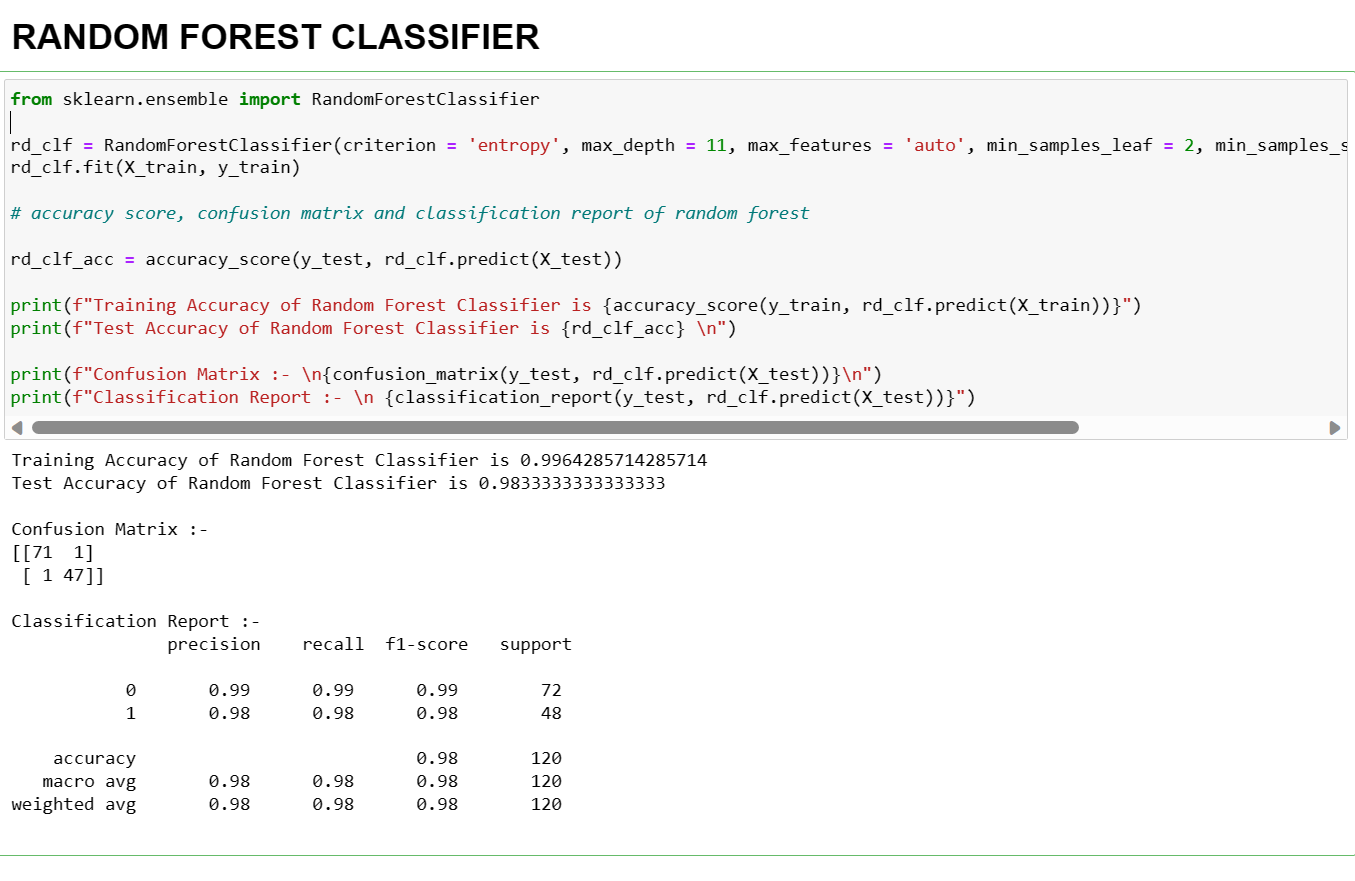
****

****

**** **** 

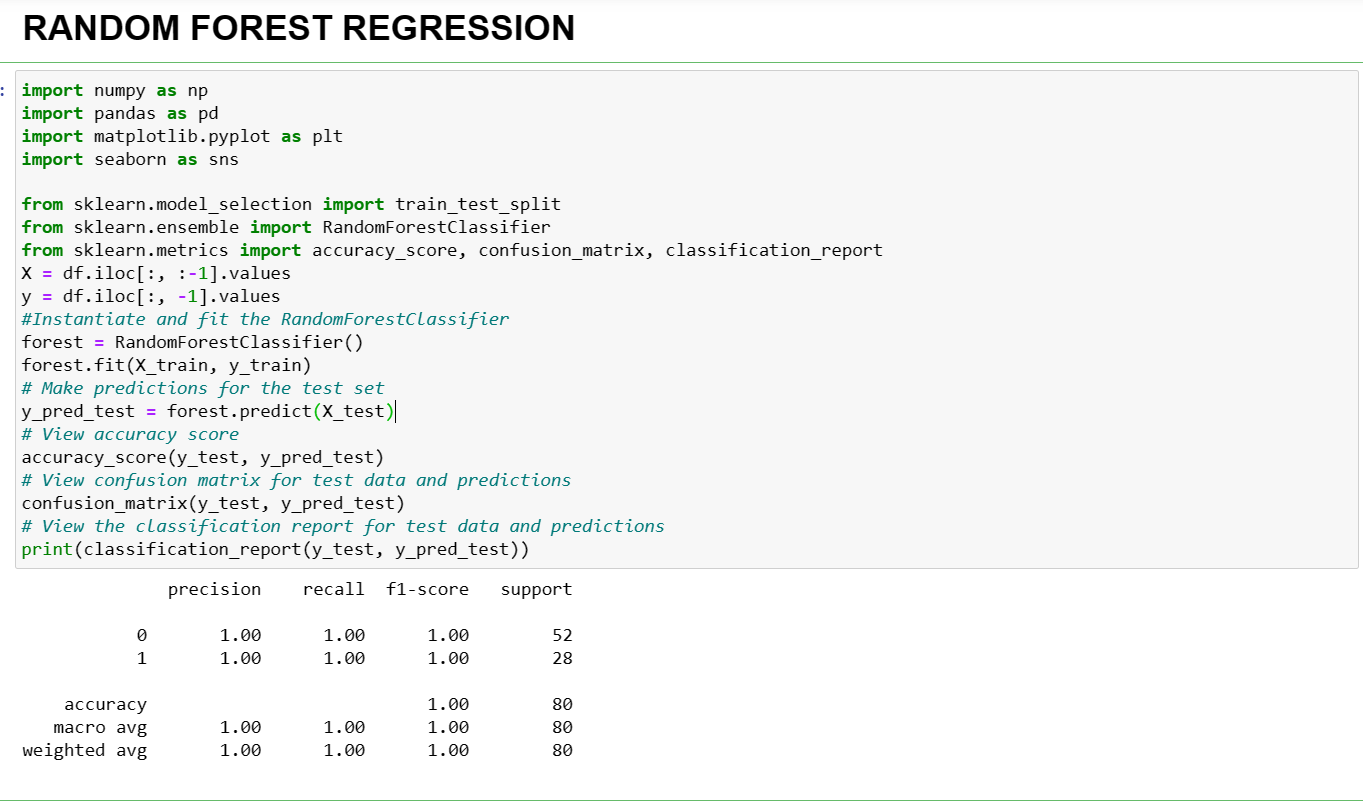
****

****

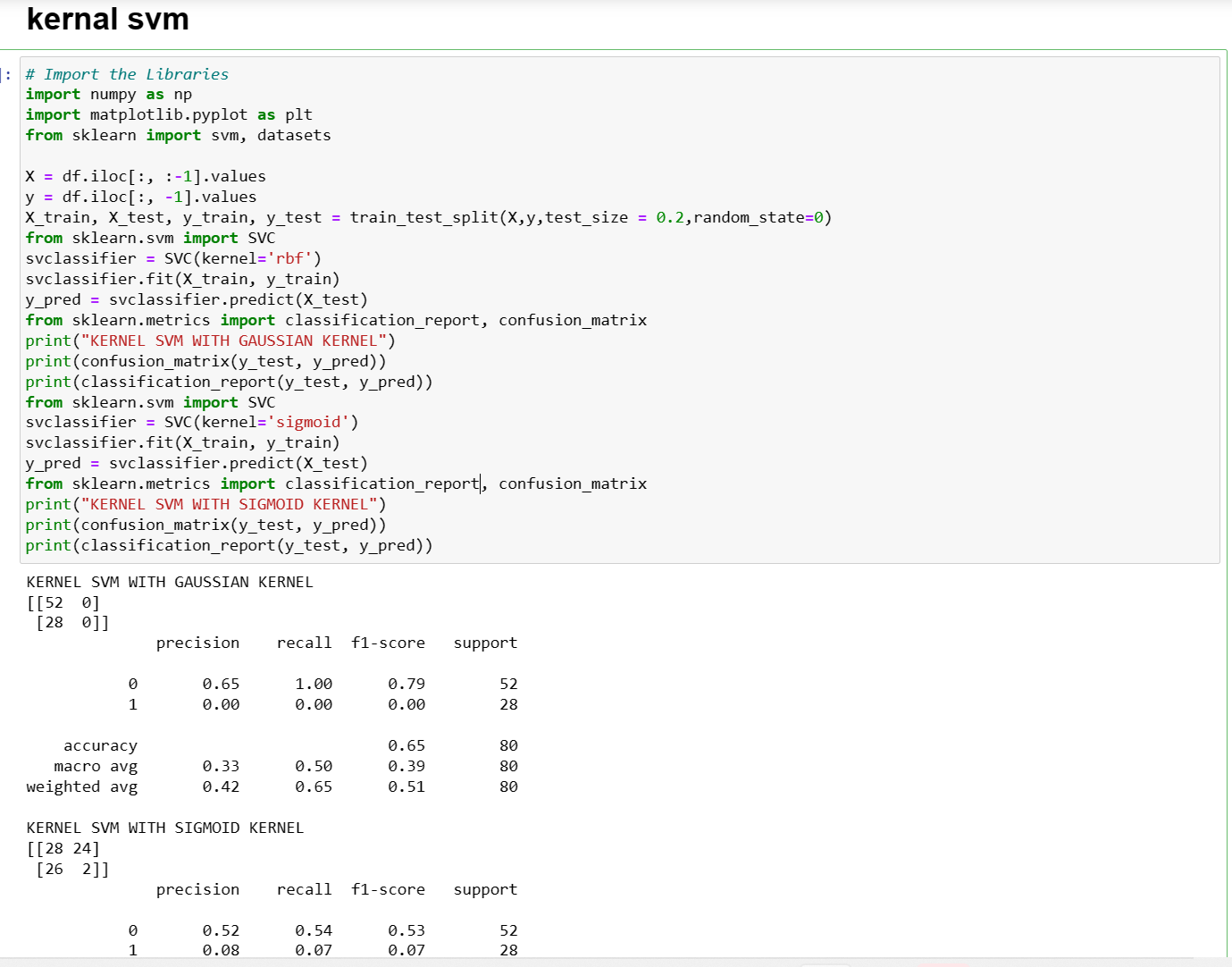
****

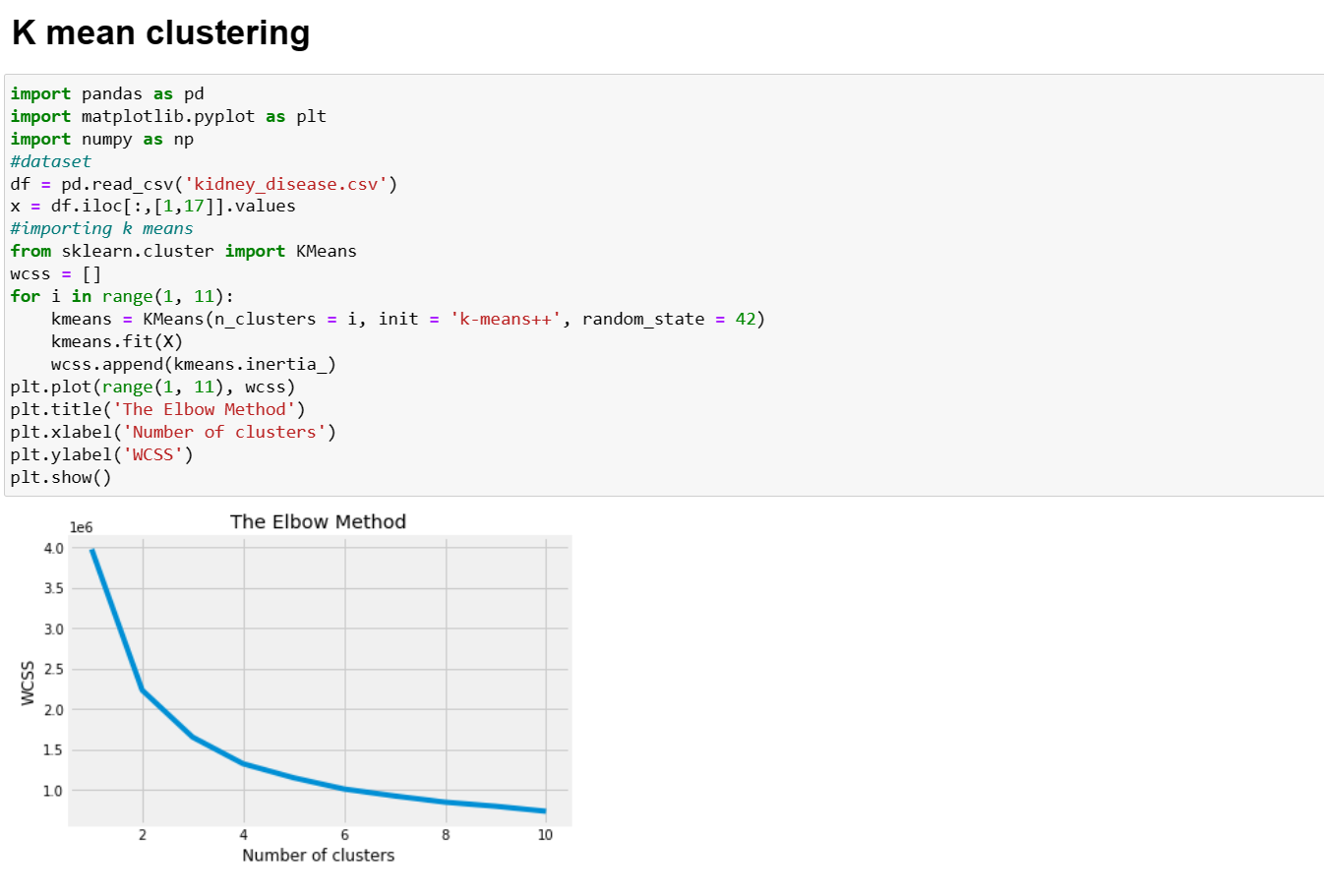
****

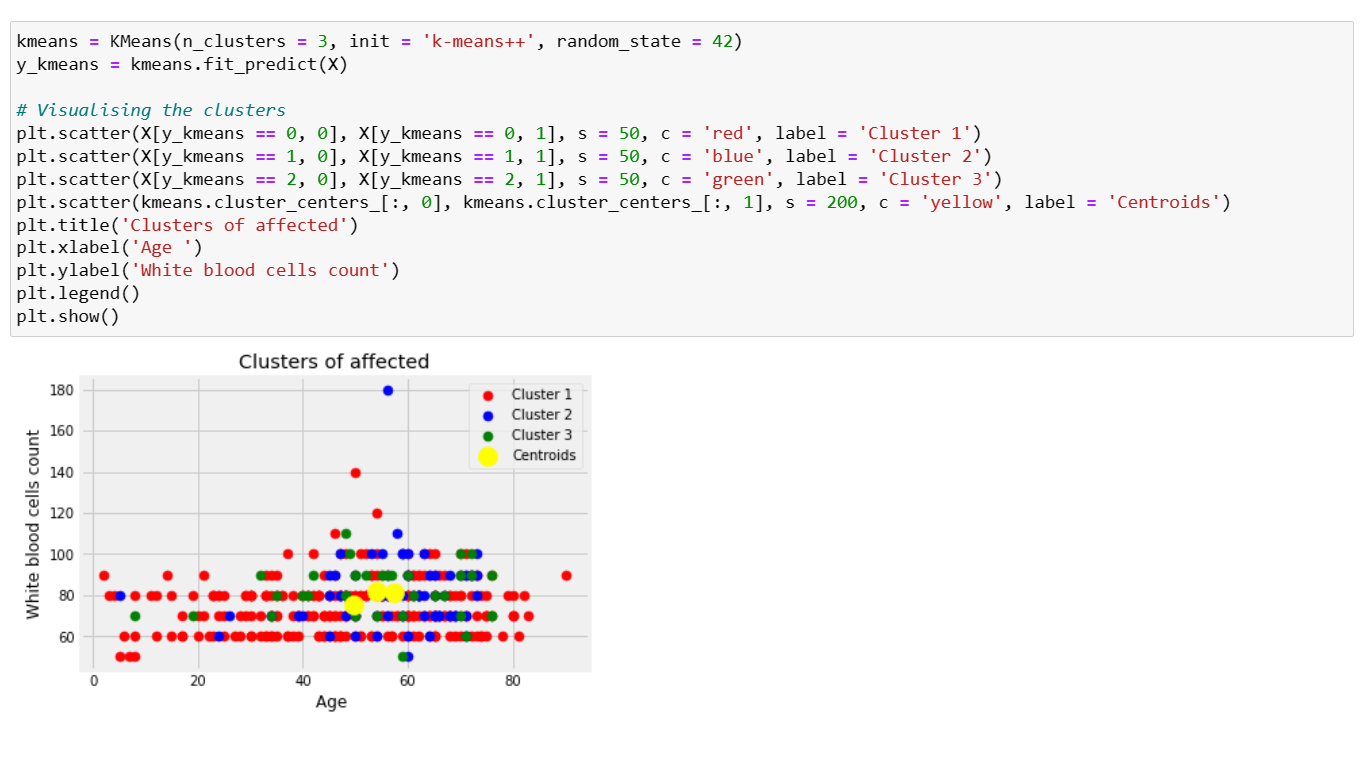
****

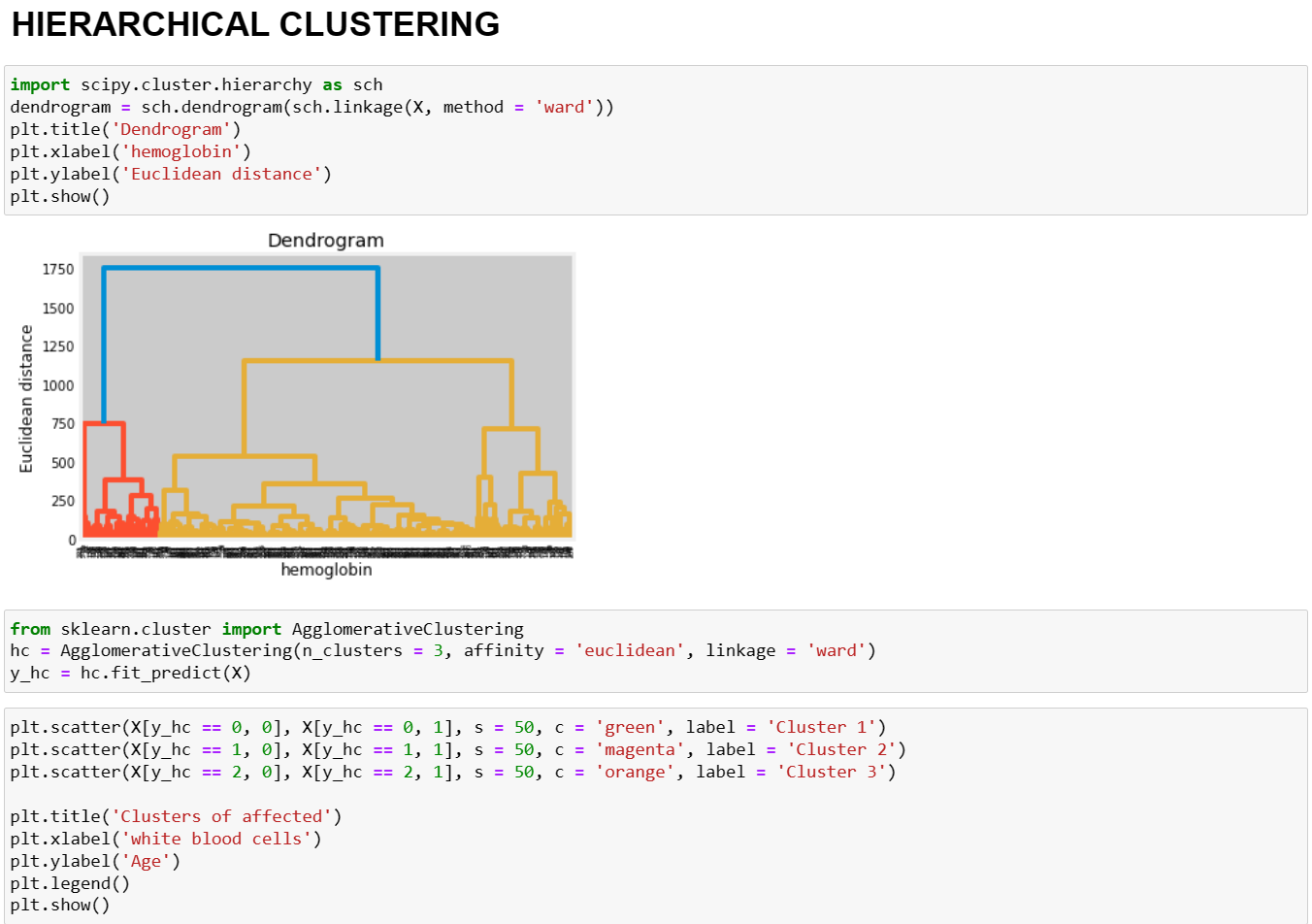
****

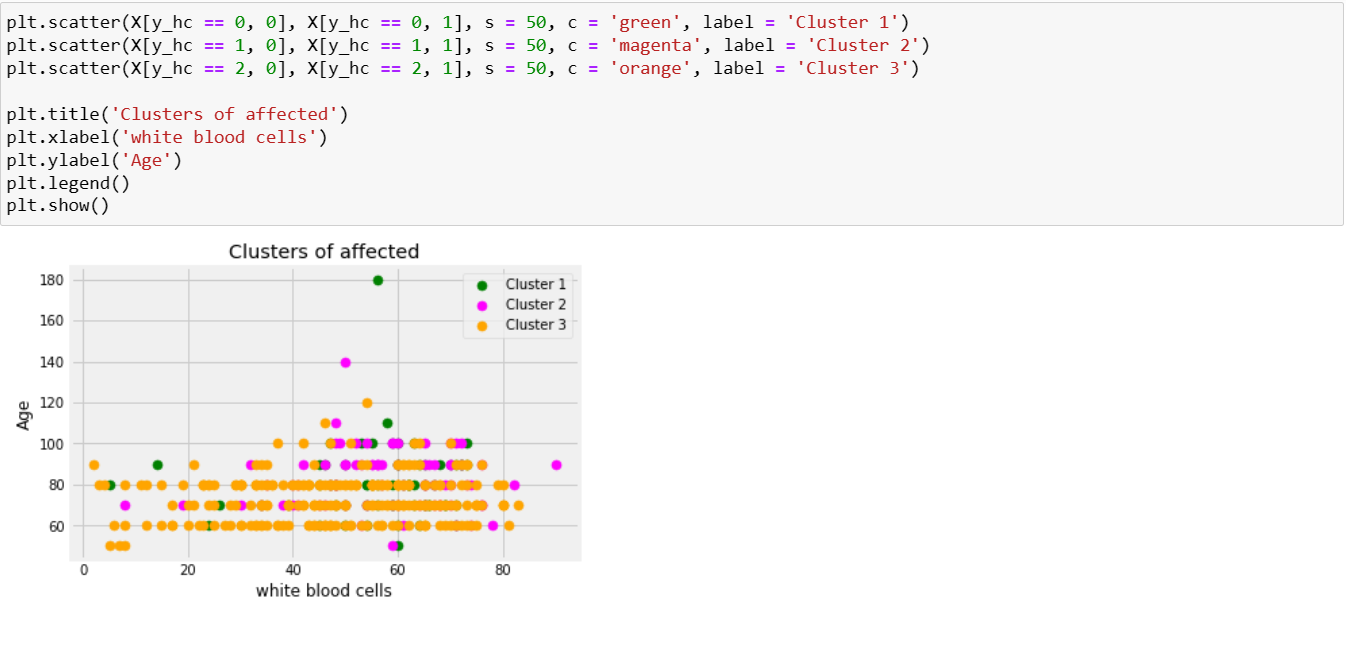
****

****

****

****

****

****

****

****

**RESULTS**

The results of each classifier have been reviewed with a variety of criteria for evaluation.

This project has utilized, data world which is open- source software for the machine learning library in Python, in several different ways. Accuracy, F1-score, precision, and recall are the evaluation metrics that were taken into consideration for this study.

According to the results, all of the models have excellent performance in terms of detecting CKD with accuracy of greater than 97% by utilizing specific gravity, and albumin, sugar, blood glucose random, serum creatinine, potassium, packed cell volume, white and red blood cell count, and diabetes mellitus characteristics, shown in. By concentrating on precision and recall, Random Forest tree 98%

**CONCLUSION**

In this research, we have used chronic kidney disease dataset collected from data world. We have developed a chronic kidney disease prediction model using all machine learning algorithms to measure the performance of the prediction model. The performance of the model depends upon various performance matrices like sensitivity, precision, recall, f1-score, confusion matrix etc. In classifications logistic regression have highest accuracy rate and in regressions random forest have highest accuracy rate.

**REFERENCES**

**Journal References:**

1. Hill N. R., Fatoba S. T., Oke J. L., Hirst J. A., O’Callaghan C. A., Lasserson D. S., et al. (2016). Global Prevalence of Chronic Kidney Disease—A Systematic Review and Meta-Analysis. PloS one, 11(7), e0158765. pmid:27383068
2. **2.**GBD Chronic Kidney Disease Collaboration (2020). Global, regional, and national burden of chronic kidney disease, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet (London, England), 395(10225), 709–733. pmid:32061315
3. **3.**Chen T. K., Knicely D. H., & Grams M. E. (2019). Chronic Kidney Disease Diagnosis and Management: A Review. JAMA, 322(13), 1294–1304. pmid:31573641
4. **4.**Levin A., Tonelli M., Bonventre J., Coresh J., Donner J. A., Fogo A. B., et al. (2017). Global kidney health 2017 and beyond: a roadmap for closing gaps in care, research, and policy. Lancet (London, England), 390(10105), 1888–1917. pmid:28434650
5. Worldwide access to treatment for end-stage kidney disease: a systematic review. Lancet (London, England), 385(9981), 1975–1982. pmid:25777665
6. **6.**Panch T, Durable E, Mattie H, Kotecha G, Celi LA, Wright M, et al. (2022) A distributed approach to the regulation of clinical AI. PLOS Digit Health 1(5): e0000040.

**Web References:**

[**https://link.springer.com/article/10.1007/s44174-022-00027-y**](https://link.springer.com/article/10.1007/s44174-022-00027-y)

[**https://www.geeksforgeeks.org/machine-learning/**](https://www.geeksforgeeks.org/machine-learning/)

**FUTURE SCOPE**

In future we will work on the Reinforcement learning and the best algorithms. By analysing historical and real-time data, doctors can predict the future situation of patients. Our method can be used to test various datasets like Heart, Cancer, and Diabetes disease. The main objective of this study was to predict the chronic kidney disease using attributes while maintaining a higher accuracy. Future research should analyse different supervised and unsupervised machine learning technique with additional performance metrics for better chronic disease prediction.

**WORKLOG**

|  |  |  |
| --- | --- | --- |
| WEEKS | DATE | **TASK DONE** |
| Week -1 | 17.02.2023 | Watched reference videos (Udemy) |
| Week -2 | 24.02.2023 | Worked out with the given dataset in python |
| Week -3 | 03.03.2023 | Read some journals and started working on our dataset. |
| Week -4 | 10.03.2023 | Done data cleaning and correlation |
| Week -5 | 17.03.2023 | Worked on data pre-processing and some models |
| Week -6 | 21.03.2023 | Worked on Classification Models |
| Week -7 | 31.03.2023 | Worked on Regression Models |
| Week -8 | 08.03.2023 | Worked on Clustering |
| Week -9 | 16.04.2023 | Comparison graph and corrections from second review |
| Week -10 | 23.04.2023 | Worked on report |