Problem Statement

Prediction of Chronic Kidney Disease using Machine Learning

Introduction:

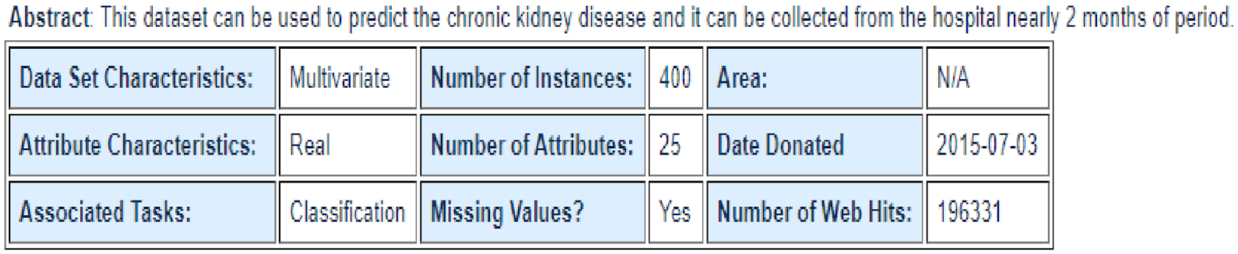
Technological development, including machine learning, has a huge impact on health through an effective analysis of various chronic diseases for more accurate diagnosis and successful treatment. Kidney disease is a major chronic disease associated with aging, hypertension, and diabetes, affecting people 60 and over. Its major cause is the malfunctioning of the kidney in disposing toxins from the blood. This study analyzes chronic kidney disease using machine learning techniques based on a chronic kidney disease (CKD).

Chronic Kidney Disease (CKD) or chronic renal disease has become a major issue with a steady growth rate. A person can only survive without kidneys for an average time of 18 days, which makes a huge demand for a kidney transplant and Dialysis.

It is important to have effective methods for early prediction of CKD. Machine learning methods are effective in CKD prediction. This work proposes a workflow to predict CKD status based on clinical data, incorporating data prepossessing, a missing value handling method with collaborative filtering and attributes selection. Out of the 11 machine learning methods considered, the extra tree classifier and random forest classifier are shown to result in the highest accuracy and minimal bias to the attributes.

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).



Abstract: This dataset can be used to predict the chronic kidney disease and it can be collected from the hospital nearly 2 months of period.

Data Set Characteristics:	Multivariate	Number of Instances:	400	Area:	N/A
Attribute Characteristics:	Real	Number of Attributes:	25	Date Donated	2015-07-03
Associated Tasks:	Classification	Missing Values?	Yes	Number of Web Hits:	196331

Data Set Information:

We use the following representation to collect the dataset

- age - age
- bp - blood pressure
- sg - specific gravity
- al - albumin
- su - sugar
- rbc - red blood cells
- pc - pus cell
- pcc - pus cell clumps
- ba - bacteria
- bgr - blood glucose random
- bu - blood urea
- sc - serum creatinine
- sod - sodium
- pot - potassium
- hemo - hemoglobin
- pcv - packed cell volume
- wc - white blood cell count
- rc - red blood cell count

- htn - hypertension
- dm - diabetes mellitus
- cad - coronary artery disease
- appet - appetite
- pe - pedal edema
- ane - anemia
- class - class

Attribute Information:

We use 24 + class = 25 (11 numeric ,14 nominal)

1.Age(numerical)

age in years

2.Blood Pressure(numerical)

bp in mm/Hg

3.Specific Gravity(nominal)

sg - (1.005,1.010,1.015,1.020,1.025)

4.Albumin(nominal)

al - (0,1,2,3,4,5)

5.Sugar(nominal)

su - (0,1,2,3,4,5)

6.Red Blood Cells(nominal)

rbc - (normal,abnormal)

7.Pus Cell (nominal)

pc - (normal,abnormal)

8.Pus Cell clumps(nominal)

pcc - (present,notpresent)

9.Bacteria(nominal)

ba - (present,notpresent)

10.Blood Glucose Random(numerical)

bgr in mgs/dl

11.Blood Urea(numerical)

bu in mgs/dl

12.Serum Creatinine(numerical)

sc in mgs/dl

13.Sodium(numerical)

sod in mEq/L

14.Potassium(numerical)

pot in mEq/L

15.Hemoglobin(numerical)

hemo in gms

16.Packed Cell Volume(numerical)

17.White Blood Cell Count(numerical)

wc in cells/cumm

18.Red Blood Cell Count(numerical)

rc in millions/cmm

19.Hypertension(nominal)

htn - (yes,no)

20.Diabetes Mellitus(nominal)

dm - (yes,no)

21.Coronary Artery Disease(nominal)

cad - (yes,no)

22.Appetite(nominal)

appet - (good,poor)

23.Pedal Edema(nominal)

pe - (yes,no)

24.Anemia(nominal)

ane - (yes,no)

25.Class (nominal)

class - (ckd,notckd)

Task to Do:

A. Missing Value Handling: In the pre-processing step, missing values have to be handled based on their distributions to achieve reasonable accuracy

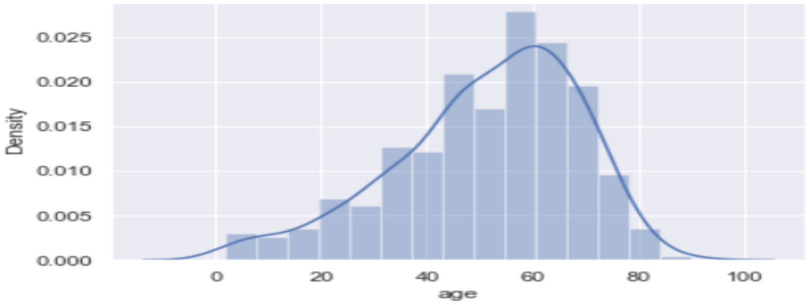
B. Data Preprocessing : Feature Selection

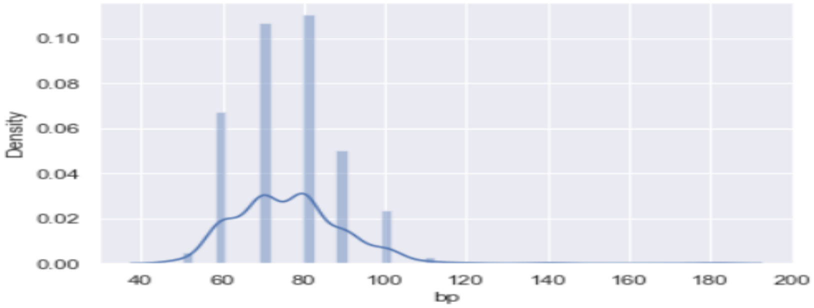
C. Model Training:In this work, 11 classification models can be taking in to consideration . They are logistic regression, k-Nearest Neighbors (KNN) regression, SVC with a linear kernel, SVC with RBF kernel, Gaussian NB, decision tree classifier, random forest classifier, XGB classifier, extra trees classifier, an ada boost classifier and a classical neural network.

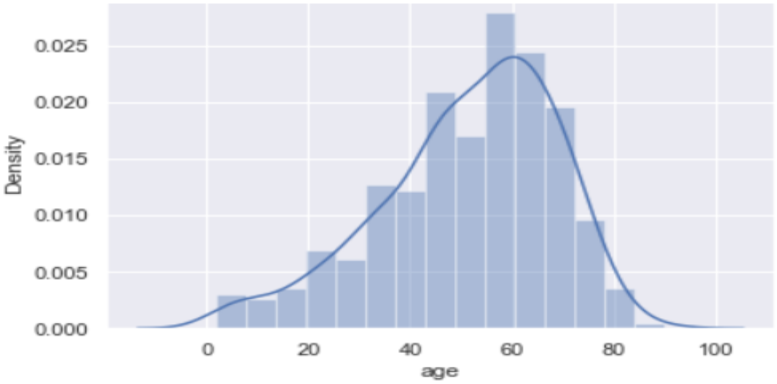
D. Find out which fits the best and gives the best Accuracy to the model.

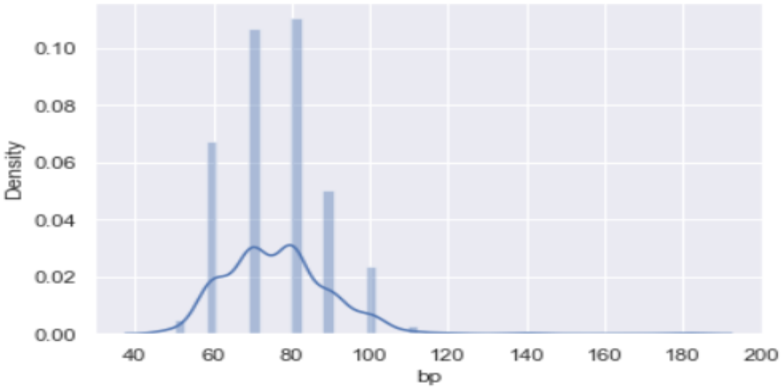
Model Evaluation:

Sample Output:

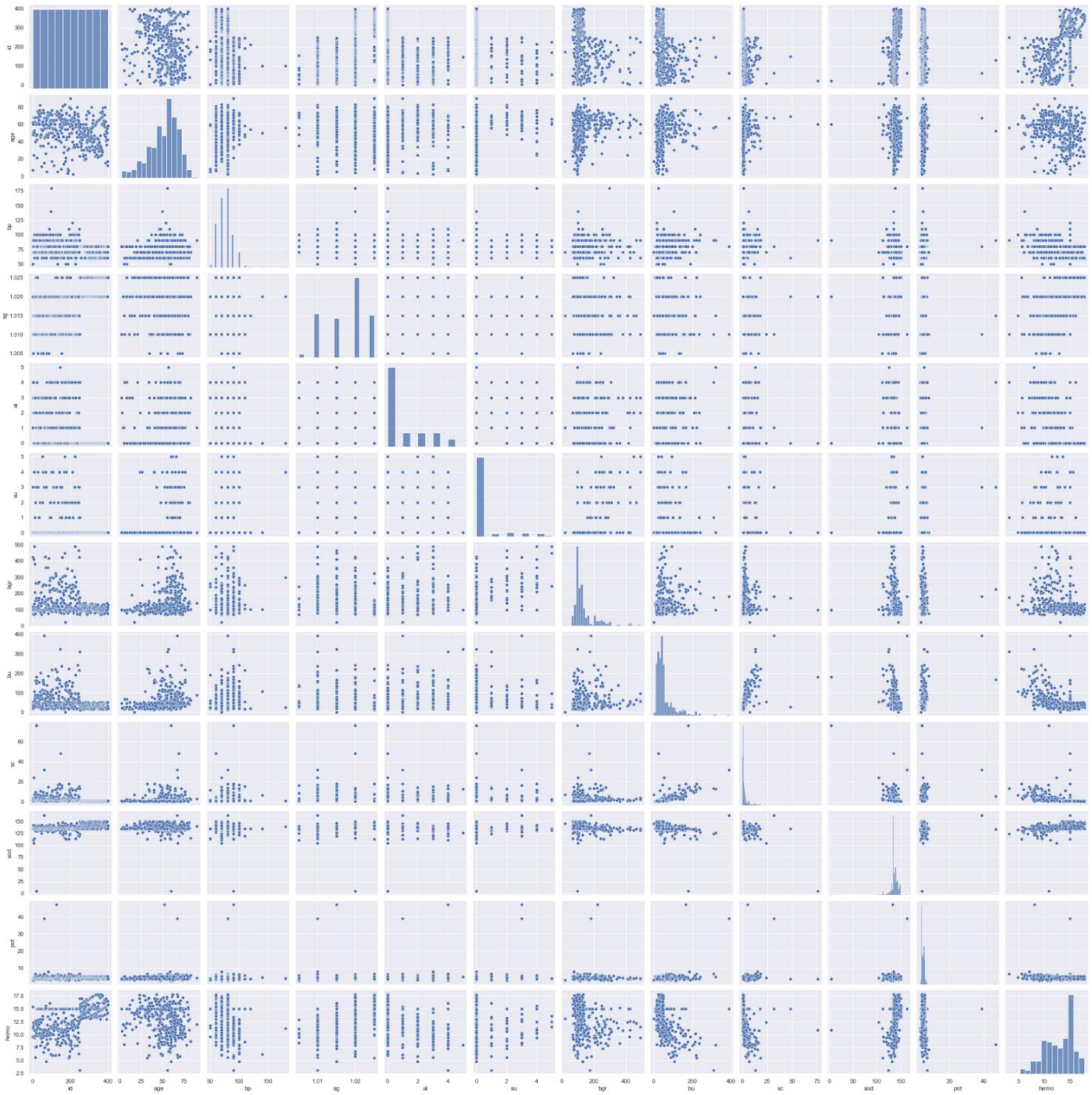




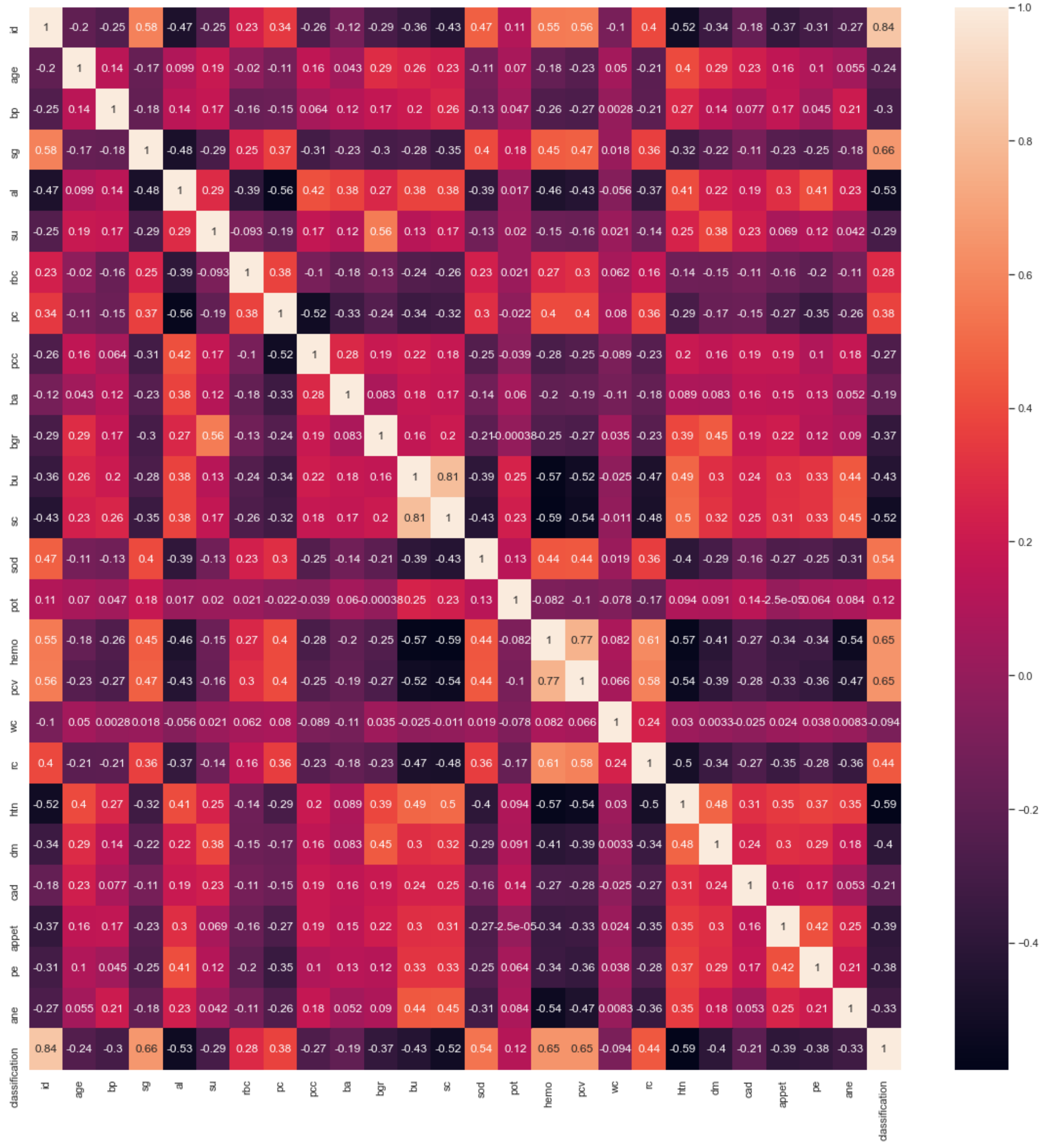




Sample Output:

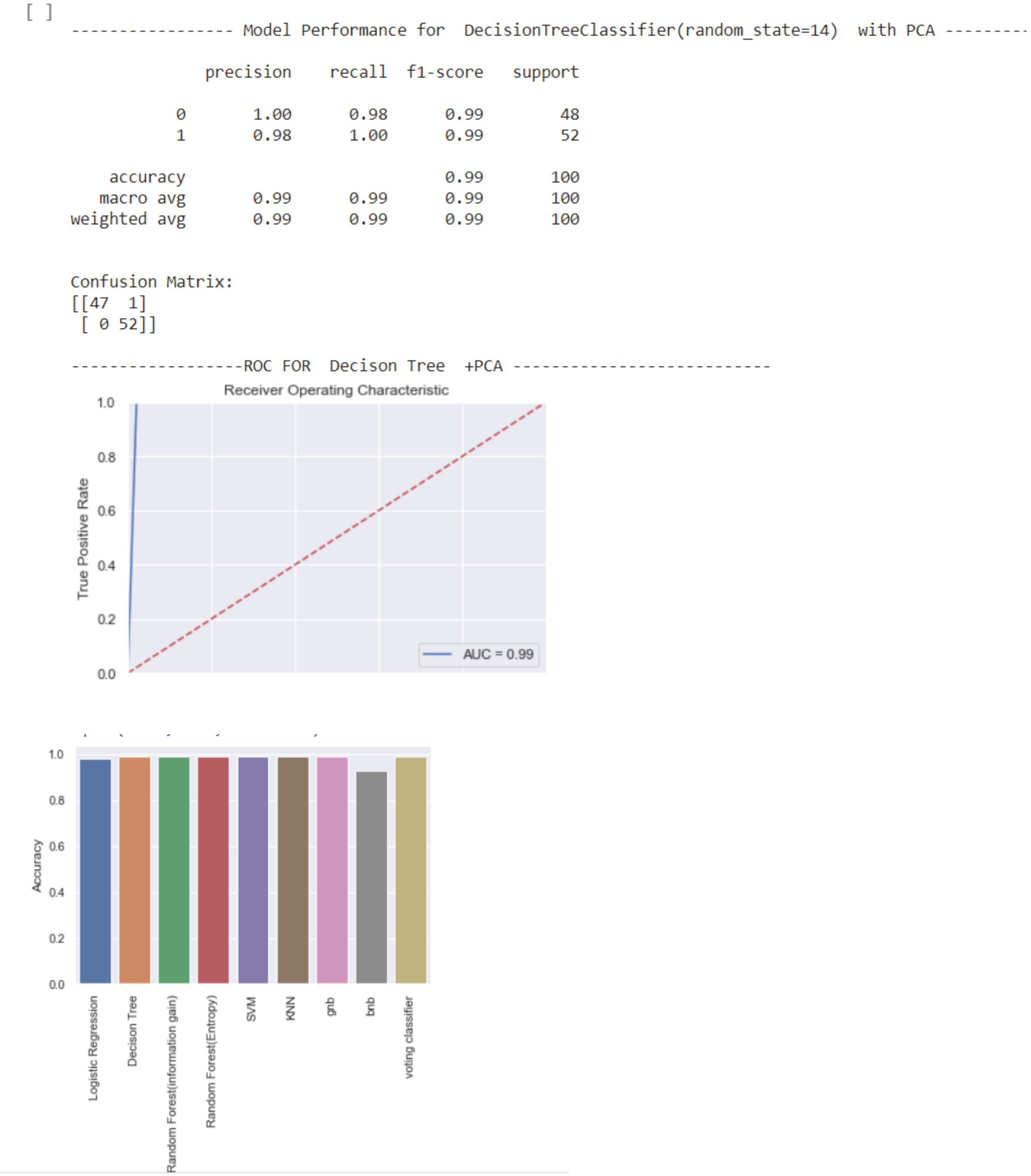


Sample Output:



	id	age	bp	sg	al	su	rbc	pc	pcc	ba	bgr	bu	sc	sod	pot	hemo	pcv	wc	rc	htn	dm	cad	appet	pe	ane	classification
id	1	-0.2	-0.25	0.58	-0.47	-0.25	0.23	0.34	-0.26	-0.12	-0.29	-0.36	-0.43	0.47	0.11	0.55	0.56	-0.1	0.4	-0.52	-0.34	-0.18	-0.37	-0.31	-0.27	0.84
age	-0.2	1	0.14	-0.17	0.099	0.19	-0.02	-0.11	0.16	0.043	0.29	0.26	0.23	-0.11	0.07	-0.18	-0.23	0.05	-0.21	0.4	0.29	0.23	0.16	0.1	0.055	-0.24
bp	-0.25	0.14	1	-0.18	0.14	0.17	-0.16	-0.15	0.064	0.12	0.17	0.2	0.26	-0.13	0.047	-0.26	-0.27	0.0028	-0.21	0.27	0.14	0.077	0.17	0.045	0.21	-0.3
sg	0.58	-0.17	-0.18	1	-0.48	-0.29	0.25	0.37	-0.31	-0.23	-0.3	-0.28	-0.35	0.4	0.18	0.45	0.47	0.018	0.36	-0.32	-0.22	-0.11	-0.23	-0.25	-0.18	0.66
al	-0.47	0.099	0.14	-0.48	1	0.29	-0.39	-0.56	0.42	0.38	0.27	0.38	0.38	-0.39	0.017	-0.46	-0.43	-0.056	-0.37	0.41	0.22	0.19	0.3	0.41	0.23	-0.53
su	-0.25	0.19	0.17	-0.29	0.29	1	-0.093	-0.19	0.17	0.12	0.56	0.13	0.17	-0.13	0.02	-0.15	-0.16	0.021	-0.14	0.25	0.38	0.23	0.069	0.12	0.042	-0.29
rbc	0.23	-0.02	-0.16	0.25	-0.39	-0.093	1	0.38	-0.1	-0.18	-0.13	-0.24	-0.26	0.23	0.021	0.27	0.3	0.062	0.16	-0.14	-0.15	-0.11	-0.16	-0.2	-0.11	0.28
pc	0.34	-0.11	-0.15	0.37	-0.56	-0.19	0.38	1	-0.52	-0.33	-0.24	-0.34	-0.32	0.3	-0.022	0.4	0.4	0.08	0.36	-0.29	-0.17	-0.15	-0.27	-0.35	-0.26	0.38
pcc	-0.26	0.16	0.064	-0.31	0.42	0.17	-0.1	-0.52	1	0.28	0.19	0.22	0.18	-0.25	-0.039	-0.28	-0.25	-0.089	-0.23	0.2	0.16	0.19	0.19	0.1	0.18	-0.27
ba	-0.12	0.043	0.12	-0.23	0.38	0.12	-0.18	-0.33	0.28	1	0.083	0.18	0.17	-0.14	0.06	-0.2	-0.19	-0.11	-0.18	0.089	0.083	0.16	0.15	0.13	0.052	-0.19
bgr	-0.29	0.29	0.17	-0.3	0.27	0.56	-0.13	-0.24	0.19	0.083	1	0.16	0.2	-0.21	-0.00038	-0.25	-0.27	0.035	-0.23	0.39	0.45	0.19	0.22	0.12	0.09	-0.37
bu	-0.36	0.26	0.2	-0.28	0.38	0.13	-0.24	-0.34	0.22	0.18	0.16	1	0.81	-0.39	0.25	-0.57	-0.52	-0.025	-0.47	0.49	0.3	0.24	0.3	0.33	0.44	-0.43
sc	-0.43	0.23	0.26	-0.35	0.38	0.17	-0.26	-0.32	0.18	0.17	0.2	0.81	1	-0.43	0.23	-0.59	-0.54	-0.011	-0.48	0.5	0.32	0.25	0.31	0.33	0.45	-0.52
sod	0.47	-0.11	-0.13	0.4	-0.39	-0.13	0.23	0.3	-0.25	-0.14	-0.21	-0.39	-0.43	1	0.13	0.44	0.44	0.019	0.36	-0.4	-0.29	-0.16	-0.27	-0.25	-0.31	0.54
pot	0.11	0.07	0.047	0.18	0.017	0.02	0.021	-0.022	-0.039	0.06	-0.00038	0.25	0.23	0.13	1	-0.082	-0.1	-0.078	-0.17	0.094	0.091	0.14	-2.5e-05	0.064	0.084	0.12
hemo	0.55	-0.18	-0.26	0.45	-0.46	-0.15	0.27	0.4	-0.28	-0.2	-0.25	-0.57	-0.59	0.44	-0.082	1	0.77	0.082	0.61	-0.57	-0.41	-0.27	-0.34	-0.34	-0.54	0.65
pcv	0.56	-0.23	-0.27	0.47	-0.43	-0.16	0.3	0.4	-0.25	-0.19	-0.27	-0.52	-0.54	0.44	-0.1	0.77	1	0.066	0.58	-0.54	-0.39	-0.28	-0.33	-0.36	-0.47	0.65
wc	-0.1	0.05	0.0028	0.018	-0.056	0.021	0.062	0.08	-0.089	-0.11	0.035	-0.025	-0.011	0.019	-0.078	0.082	0.066	1	0.24	0.03	0.0033	-0.025	0.024	0.038	0.0083	-0.094
rc	0.4	-0.21	-0.21	0.36	-0.37	-0.14	0.16	0.36	-0.23	-0.18	-0.23	-0.47	-0.48	0.36	-0.17	0.61	0.58	0.24	1	-0.5	-0.34	-0.27	-0.35	-0.28	-0.36	0.44
htn	-0.52	0.4	0.27	-0.32	0.41	0.25	-0.14	-0.29	0.2	0.089	0.39	0.49	0.5	-0.4	0.094	-0.57	-0.54	0.03	-0.5	1	0.48	0.31	0.35	0.37	0.35	-0.59
dm	-0.34	0.29	0.14	-0.22	0.22	0.38	-0.15	-0.17	0.16	0.083	0.45	0.3	0.32	-0.29	0.091	-0.41	-0.39	0.0033	-0.34	0.48	1	0.24	0.3	0.29	0.18	-0.4
cad	-0.18	0.23	0.077	-0.11	0.19	0.23	-0.11	-0.15	0.19	0.16	0.19	0.24	0.25	-0.16	0.14	-0.27	-0.28	-0.025	-0.27	0.31	0.24	1	0.16	0.17	0.053	-0.21
appet	-0.37	0.16	0.17	-0.23	0.3	0.069	-0.16	-0.27	0.19	0.15	0.22	0.3	0.31	-0.27	-2.5e-05	-0.34	-0.33	0.024	-0.35	0.35	0.3	0.16	1	0.42	0.25	-0.39
pe	-0.31	0.1	0.045	-0.25	0.41	0.12	-0.2	-0.35	0.1	0.13	0.12	0.33	0.33	-0.25	0.064	-0.34	-0.36	0.038	-0.28	0.37	0.29	0.17	0.42	1	0.21	-0.38
ane	-0.27	0.055	0.21	-0.18	0.23	0.042	-0.11	-0.26	0.18	0.052	0.09	0.44	0.45	-0.31	0.084	-0.54	-0.47	0.0083	-0.36	0.35	0.18	0.053	0.25	0.21	1	-0.33
classification	0.84	-0.24	-0.3	0.66	-0.53	-0.29	0.28	0.38	-0.27	-0.19	-0.37	-0.43	-0.52	0.54	0.12	0.65	0.65	-0.094	0.44	-0.59	-0.4	-0.21	-0.39	-0.38	-0.33	1

Sample Output:



```
[ ]
     ---------------- Model Performance for  DecisionTreeClassifier(random_state=14)  with PCA ---------

                   precision    recall  f1-score   support

                0       1.00      0.98      0.99        48
                1       0.98      1.00      0.99        52

         accuracy                           0.99       100
        macro avg       0.99      0.99      0.99       100
     weighted avg       0.99      0.99      0.99       100


     Confusion Matrix:
     [[47  1]
      [ 0 52]]

     -----------------ROC FOR  Decison Tree  +PCA --------------------------
```