**Lab Assignment 1**

Arundhati Das

BT17GCS016

Now write code to:

1. Read the data
2. Check if there is any missing data
3. Select any 5 numeric features
4. For each pair of the selected features (i.e. 10 pairs in all)

Build a scatter plot of the numeric values. Use separate colors to represent the samples belonging to the two classes (ckd / notckd). In a given plot you will have one point per patient

1. Build a report (word file) that contains the plots. Now analyze these plots and write down if you can observe any pattern / trend.

**Dataset Used:** Chronic Kidney Disease Dataset from <https://archive.ics.uci.edu/ml/datasets.php>

**Code:**

import pandas as pd

import matplotlib.pyplot as plt

df=pd.read\_csv("kidney\_disease.csv")

df.head()

import matplotlib.pyplot as plt

age=df["'age'"]

bp=df["'bp'"]

ckd=df["'class'"]

colors=[]

for i in ckd:

  if i=="ckd":

    colors.append('red')

  else:

    colors.append('green')

plt.scatter(age,bp,c=colors)

plt.xlabel('Age')

plt.ylabel('Blood Pressure')

plt.show()



import matplotlib.pyplot as plt

bgr=df["'bgr'"]

bu=df["'bu'"]

ckd=df["'class'"]

colors=[]

for i in ckd:

  if i=="ckd":

    colors.append('red')

  else:

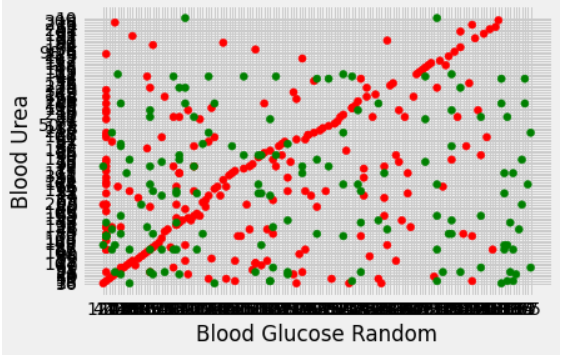
    colors.append('green')

plt.scatter(bgr,bu,c=colors)

plt.xlabel('Blood Glucose Random')

plt.ylabel('Blood Urea')

plt.show()



import matplotlib.pyplot as plt

sc=df["'sc'"]

sod=df["'sod'"]

ckd=df["'class'"]

colors=[]

for i in ckd:

  if i=="ckd":

    colors.append('red')

  else:

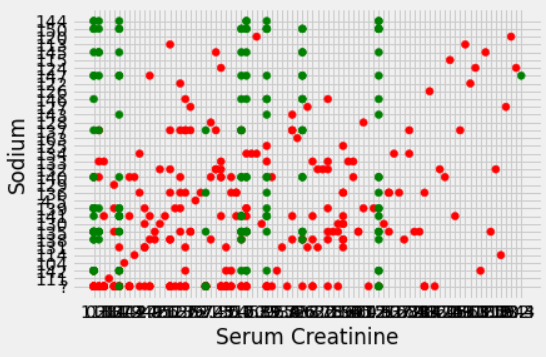
    colors.append('green')

plt.scatter(sc,sod,c=colors)

plt.xlabel('Serum Creatinine')

plt.ylabel('Sodium')

plt.show()



import matplotlib.pyplot as plt

pot=df["'pot'"]

hemo=df["'hemo'"]

ckd=df["'class'"]

colors=[]

for i in ckd:

  if i=="ckd":

    colors.append('red')

  else:

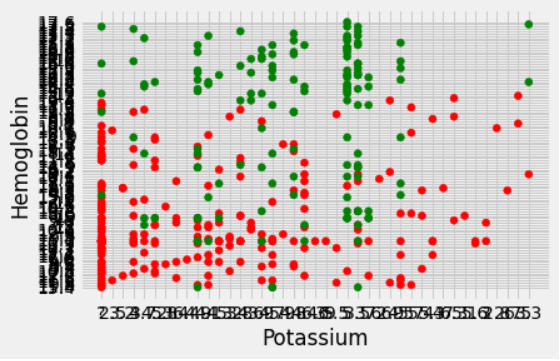
    colors.append('green')

plt.scatter(pot,hemo,c=colors)

plt.xlabel('Potassium')

plt.ylabel('Hemoglobin')

plt.show()



import matplotlib.pyplot as plt

pcv=df["'pcv'"]

wbcc=df["'wbcc'"]

ckd=df["'class'"]

colors=[]

for i in ckd:

  if i=="ckd":

    colors.append('red')

  else:

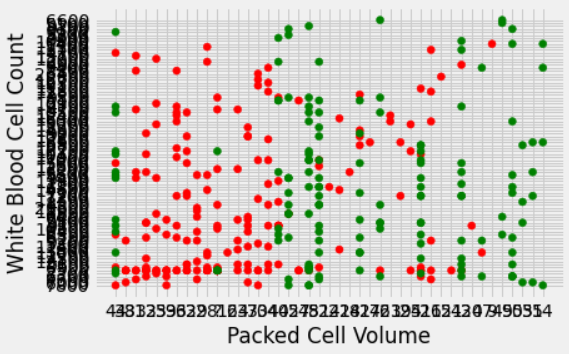
    colors.append('green')

plt.scatter(pcv,wbcc,c=colors)

plt.xlabel('Packed Cell Volume')

plt.ylabel('White Blood Cell Count')

plt.show()



**Analysis:**

A scatter plot is used to visualize the relation between the two numeric features. Red dots are patients with Chronic Kidney Disease (CKD) whereas Green dots are patients without CKD.

I made 5 scatter plots by selecting the following numeric features :

* **Age & Blood Pressure:** The middle part of the plot mostly shows the patients having CKD whereas the lower-right side of the plot mostly shows the patients not having CKD.
* **Blood Glucose Random & Blood Urea:** Mostly the patients having CKD are shown in a linear fashion (i.e a straight line ) whereas the right most part of the plot mostly shows the patients not having CKD.
* **Serum Creatinine & Sodium:** The lower part of the plot mostly shows the patients having CKD whereas the upper part of the plot mostly shows the patients not having CKD.
* **Potassium & Hemoglobin:** The lower part of the plot mostly shows the patients having CKD whereas the upper part of the plot mostly shows the patients not having CKD.
* **White Blood Cell Count & Packed Cell Volume:** Left side of the plot mostly shows the patients having CKD whereas the right side of the plot mostly shows the patients not having CKD.