**Kidney Dataset**

**Features**

1. Age- age
2. Bp- blood pressure
3. Sg- specific gravity
4. Al- albium
5. Su- sugar
6. Rbc- red blood cells
7. Pc- pus cells
8. Pcc- pus cell clumps
9. Ba- bacteria
10. Bgr- blood glucose random
11. Bu- blood urea
12. Sc- serum creatinine
13. Sod- sodium
14. Pot- potassium
15. Hemo- haemoglobin
16. Pcv- packed cell volume
17. Wc- wbc count
18. Rc- rbc count
19. Htn- hypertension
20. Dm- diabetes mellitus
21. Cad- coronary artery disease
22. Appet- appetite
23. Ane- anaemia
24. Classification- class

**Renaming columns to make it user-friendly**

So, I created a data description file and imported in the existing dataset and resettled the index.

**Type conversion**

RBC count, packed cell volume and WBC count are object type. We need to convert that into numerical type.

**Dropping ID column**

Dropping ID column as it is the unique identifier for each row and it won’t help us find any insights from the data.

**Checked Numerical & Categorical Data**

11 categorical data classes & 14 numerical data classes

Checked the dirtiness in categorical data. There were some incorrect values so we need to replace those. After replacing the values, we have now on NAN values.

**Feature Distribution**

1. Age looks like it is left skewed
2. Blood glucose random is right skewed
3. Blood Urea is also a bit right skewed
4. Rest of the features are lightly skewed

**Label Distribution**

Few features have imbalance categories

**Correlation**

Positive Correlation:

1. Specific gravity -> Red blood cell count, Packed cell volume and Hemoglobin
2. Sugar -> Blood glucose random
3. Blood Urea -> Serum creatinine
4. Hemoglobin -> Red Blood cell count <- packed cell volume

Negative Correlation:

1. Albumin, Blood urea -> Red blood cell count, packed cell volume, Hemoglobin
2. Serum creatinine -> Sodium

**Positive Correlation**

Both distributions are quite different, distribution CKD is quite normal and evenly distributed but not CKD distribution is a little bit left-skewed but quite close to a normal distribution.

**Analyzed RBC Distribution**

1. RBC count range is from 2 to <4.5 and Hemoglobin between 3 to <13 is mostly classified as positive for chronic kidney disease (i.e ckd).
2. RBC count range >4.5 to ~6.1 and Hemoglobin between >13 to 17.8 are classified as negative for chronic kidney disease (i.e nockd).

**Negative Correlation**

Albium has 0 effect on kidney disease

**Data Cleaning**

Filling missing values with random values

\*\*Further information are given in python file named as Kidney Disease Prediction