**Lab Instructions – 1**

Data plays a central role in Machine learning. So, it is essential that you learn to ***look*** at data. This exercise will introduce you to data:

Download the “Chronic\_Kidney\_Disease Data Set” from <https://archive.ics.uci.edu/ml/datasets.php>

Read the description very carefully and make sure that you understand it. The following segment is particularly relevant:

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)

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.
2. Upload the report.

Obviously, if you don’t make any mistakes, then all of you will get the same plot if you choose the same subset of numeric features. It is your analysis (in step 5) that is important. *Write the analysis in your own words.*

By the way, this a real world system. Doctors need these systems to predict if any patient is at high risk of developing kidney disease. Obviously, such patients would be advised accordingly.