**Data-Science Task -** Machine-Learning problem: Predicting CKD (chronic kidney disease) based on subject health records

The goal of this task is to predict the subject medical condition (CKD / healthy).

**Data-set**

The CKD is anonymized simulated dataset with clinical measures acquired from different sites in the world. Each row in the dataset corresponds to a single subject. Pay attention that as usually happens in real world problems there are some missing values and potentially some outliers in the dataset.

CKD.csv file in (ML\_problem folder) including 24 health related attributes, denoted as features (e.g. age, red blood cell, white blood cell count, etc.) for 400 subjects participated in the experiment.

* 1. **Data preparation**
     1. Randomly split the data into training and test set
     2. Handle missing data (features/clinical data) by imputing them
     3. Scale the features according to a standard scaling method
  2. **Data exploration:**
     1. Summarize the input data in terms of N, average, % of missing data, percentiles.
     2. Plot distribution of 2-3 features that you’d like to explore
     3. Detect outliers in the data according to any outliers- detection method you know, and think is appropriate to the data.
  3. **Modeling**
     1. Perform feature-selection to extract the most important features for the classification.
     2. Perform multi-feature classification using at least two classification techniques. Visualize the results.
  4. **Appendix- 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)