**Data-Science Task**

The following task consists from typical data-science workflow steps. You will be instructed what to do in each step and implement a Python code where needed.

You will get two anonymized simulated datasets 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.

The EEG-dataset will be used for solving typical data management problems.

The CSK dataset will be used for solving a classification problem.

The focus of the task is on the ability to apply typical Data-Science work flow programmatically, and less on the actual prediction results. (But of course, if you manage to get to a model with high accuracy it's preferable).

You have 24 hours to complete the task and send the results back.

Good-luck!

1. **Part A: Data management problem**

Use the files in the DM\_problem folder for this task.

* 1. The file df\_eeg.xlsx holds the EEG features (column J-end) with some meta data (columns A-I) and file df\_clinical.xlsx holds the clinical data.

Read both tables into pandas data frames. Remarks about EEG table:

* Note that not all EEG features are numeric.
* 'subject|blinded\_reference' contains the subjects' number
* 'taskData|visit' contains the visit number
* 'taskData|recordingDate' contains the recording date
* Note that most subjects have more than one visit
* 'subject|group' columns indicate Healthy/MDD for the different groups
  1. Merge the two tables based on subject number ('subject|blinded\_reference) and recording date ('taskData|recordingDate').
  2. Handle duplicate records in the dataset by removing them. How many duplicates were in the eeg dataset?
  3. Write a SQL query to count the number of subjects per visit type for EGI64 recording system
  4. Summarize the dataset in terms of N per group, mean and std per feature.  
     You can use a pivot-table to summarize the results.

1. **Part B- 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). The entire data-set (Attached to this document) should be divided into training and test groups (where 60% are used for training).

**Data-set**

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. Use at least 2 types of known feature-selection techniques.
     2. For each of the selected features, build single-features classifier. Visualize the single-feature classification properties using the ROC curve (include AUC, accuracy, sensitivity and specificity measures) and show the correlation between all selected features. Use at least two classification technique for this task. Please give a brief explanation on each technique and how it differs from other techniques.
     3. Perform multi-feature classification using at least two classification techniques. Visualize the results. Please give a brief explanation on each technique and how it differs from other techniques.
  4. **Reporting**
     1. Summarize the results in a short pptx
     2. If you had more time, what would you suggest doing in order to improve the prediction accuracy?
  5. **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)