Practical Machine Learning

*Instructions:*

* *Use of your class notebook as a reference is allowed*
* *Create just one .py / one .pynb file for the entire test. You can insert comments wherever needed in that file. Also create one .docx(word) file(if required) for your solutions copy pasted from py console*
* *Upload all file(s) as the output for the test*
* *Do 5-fold cross-validation in all the Machine Learning problems with random\_state = 2021*

1. Consider the dataset named *HR\_comma\_sep.csv (10 marks)*

The outcome class is contained in a factor variable called **Left**

Variable **Type** is the target(response) variable. Try the following algorithms and examine which of these is the best fit for this data (with log loss):

1. X G Boost Classifier (learning rate=[0.001,0.5], trees=[20,30], max\_depth=[3,5] ) ( 4 marks )
2. Logistic Regression ( K-Fold CV ) ( 2 marks )
3. Random Forest ( max\_depth=[None, 4,7], max\_features=[3,5,6] ) ( 4 marks )
4. Consider the dataset *insurance.csv (10 marks)*

It is a data of the medical **charges** and various characteristics of corresponding to every customer. Build a model with **charges** as response variable and other variables as features.

Try the following algorithms and examine which of these is the best fit for this data based on **Mean Square Error**:

1. Lasso (alpha = [0, 0.1, 1, 1.5, 2] with Grid Search CV ( 5 marks)
2. X G Boost Regressor (learning rate=[0.001,0.5,0.3, 0.6], trees=[20,30], max\_depth=[3,4,5] ) ( 5 marks )
3. Consider the dataset *nutrient.csv*  *(10 marks)*

Ignore **Food\_Item** variable or set it in index.

1. Carry out a cluster analysis of the data. Do the K-means clustering and also choose the appropriate number of clusters (try for [3,4,5,6,7,8,9,10] clusters) based on criterion of Silhouette Scores. Generate a line plot to examine the Silhouette Scores. (7 marks)
2. Carry out PCA and generate the number of principal components which capture 70% variation at least (3 marks )
3. Consider dataset *ZILLOW-M1301\_MLPSF.csv*. Try out  *(10 marks)*

Split the data into train and test with test data as the last 5 values.

Try the following on this dataset:

1. All Smoothing Methods (5 marks)
2. Pmdarima package function auto\_arima( ) (5 marks)

Also evaluate both models with Mean Squared Error.