Practical Machine Learning

1. Consider the dataset named *Glass.csv (10 marks)*

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

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. Quadratic Discriminant Analysis ( K-Fold CV ) ( 3 marks )
2. Logistic Regression ( K-Fold CV ) ( 3 marks )
3. Gradient Boosting ( learning\_rate:[0.01,0.1,0.5], n\_estimators:[20,50], max\_depth: [2,3,None]) ( 4 marks )
4. Consider the dataset *Sacremento.csv (15 marks)*

It is a data of the **price** and various characteristics of corresponding real estate property. Build a model with price as response variable and other variables as features. Ignore the variable zip and city.

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

1. Ridge (alpha = [0, 0.1, 1, 1.5, 2] with Grid Search CV ( 5 marks )
2. Decision Tree ( max\_depth=[3,5], min\_samples\_split=[4,10], min\_samples\_leaf=[2,5] ) with Grid Search CV

( 5 marks )

1. Random Forest ( max\_features=[4,6,9] ) with Grid Search CV

( 5 marks )

1. Consider the dataset *USArrests.csv*  *(10 marks)*

Ignore **ID** variable.

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 suggest the number of principal components which capture 70% variation at least. (3 marks )
3. Consider dataset *BUNDESBANK-BBK01\_WT5511.csv*. Try out  *(5 marks)*

Split the data into train and test with test data as the last 10 values.

1. Holt’s Linear Trend (2 marks)
2. Holt-Winter’s Method with additive seasonality (3 marks)

Also evaluate both models with Mean Squared Error.