End of the semester project – CES550 Spring 2021

The project is due on May 15, 2021

1- Objective

To build a meaningful classification model on the given dataset.

2- Dependencies

Python > 2.7
Scikit Learn package
Pandas toolkit (If it is needed)
Numpy toolkit (If it is needed)

3- Dataset

For this assignment we use **Breast Cancer Wisconsin (Diagnostic) Dataset**. The data can be downloaded from the UC Irvine Machine Learning Repository.

The first dataset looks at the predictor classes:

- Malignant or
- Benign breast mass

The features characterize cell nucleus properties and were generated from image analysis of fine needle aspirates (FNA) of breast masses

- Sample ID (code number)
- Clump thickness
- Uniformity of cell size
- Uniformity of cell shape
- Marginal adhesion
- Single epithelial cell size
- Number of bare nuclei
- Bland chromatin
- Number of normal nuclei
- Mitosis
- Classes, i.e. diagnosis

4- Tasks

a. Data Analysis and missing data analysis.
   i. Is there missing data?
   ii. Can we afford to remove data points?
   iii. Do we use imputation (and introduce additional uncertainty)?

b. Features Engineering
   i. Features distribution plot (for all features) (Figure1)

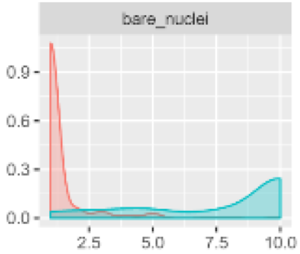

Figure1: A plot of density distribution a field. The red plot is for benign and blue is for malignant categories.

ii. Scaling
iii. Imputation
iv. Handling Outliers

c. Feature Analysis
   i. Correlation Analysis
   Generate a heap map plot for all features (Figure 2)

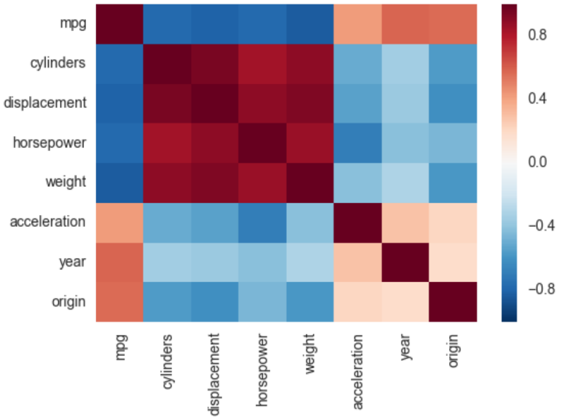

Figure2: An example of heat map for feature correlations

   ii. Univariate Selection in Scikit
   SelectKBest class from Scikit can be used to choose n best features.

d. Principal Component Analysis (PCA)
Run PCA and plot the PC1 and PC2 for two categories

e. Training, Validation, Testing
You can divide your data set as explained in previous assignment.

f. After you perform all above, select your models. You need to choose two models one from models that we studied in the class and the second one should be a new model. You can consider **ANN** as a new model since we did not have any project on it before.

i. Decision Tree base model
ii. Boosted Tree
iii. Random Forest
iv. SVM