**ML Assignment**

**Name:**

**Student Number:**

**Note: Read the following carefully**

* Any late submission will be considered as a grade of 0%

### Only one submission is allowed. (one for code )

### All the tasks must be written in Python.

### Upload only your code. Name It as B2GX(X is your group number,for example group 1 needs to name their file as B2G1 )  with your code.

### Use Jupyter Notebook.

### DO NOT MENTION ANY NAME IN THE FILE

### I do not accept your report or code through email.

### Rubrics:

### Task 1: 12.5 Marks

### Task 2: 12.5 Marks

### Note: Please review the attached example before you start this assignment .This example will prepare you and give idea about this assignment.

**Task 1: Unsupervised Learning -Clustering -Week 3 Content**

### Objective: Use Kmeans clustering to identify dominant color in the image.

### 1. import the required libraries

### 2. set no. of clusters to 3

### 3. load the rgb image and use reshape to convert it into a list of pixels

### Note:- As the image is of RGB type, each pixels will have 3 values.

### [Hint:Use cv2 module]

### 4. Use KMeans to cluster the list of pixels into 3 clusters

### [Hint: Refer to KMeans module of sklearn.cluster]

### 5. Predict the cluster label of every pixel in the image and plot it back as an image.

### 6. Find out the three dominant color in the image.

**Task 2 -** **Supervised Learning -SVC Comparison and effects/benefits of Normalization/Standardization -Week 3 content**

**Part 1: Data Loading, Feature Selection, Cleaning, preprocessing to construct an input from Data Source**

**1. Load data.csv data file using pandas.**

**2. Print out some sample rows and inspect the data sample to identify the dependent variables and independent variables.**

**3. Examine the values of each attribute and select a set of attributes only that would affect to predict dependent variable.**

**4. Split data between dependent (Y) and independent variables (X).**

**5. Determine the data type (Discrete, Continuous, Categorical) of each attribute.**

**& Encode the dependent variable (as it is a string datatype) using a proper encoding scheme.**

**Part 2: Scaling, Modeling and Hyper-parameter tuning.**

**6. Split the data into train and test data, with a ratio of 70% train and 30% test.**

**Don't forget to set a random state to maintain experiment consistency.**

**7. Demonstrate results before and after using Standard Scaler.**

**Note Observations on correct way to use fit and transform methods on train and test data.**

**Should you fit any preprocessing algorithms on the whole dataset? Why?**

**8. Also output classifier train and test accuracy scores and confusion matrix for each result.**

**9. Finally perform Hyper-parameter tuning using GridSearchCV or RandomSearchCV,**

**and note down the result train and test accuracies as well as the best parameters.**

**Write Observations on benefits/disadvantages of Standardization/Normalization.**