AUA, Machine Learning
Final Exam

Check your AUA ID = ******XY and take the last 2 numbers. Use them as the value of random state parameter in the functions that simulate data. We recommend using the same random_state value in all functions across your notebook when such parameter exists.

$$X\_1, y\_1 = make\_classification(n\_samples = 5000, n\_features = 5, n\_informative = 3, n\_clusters\_per\_class = 1, random\_state = XY, class\_sep = 2, flip\_y = 0.1, n\_classes = 5)$$

***

Problem 1.1. score = 10

Use **train_test_split** function and split $\boldsymbol{X_1, y_1}$ into 70% - 30% portions, respectively. Apply **DecisionTreeClassifier** to the **Train Data** and tune parameter **ccp_alpha (cost complexity pruning)** to maximize the **Accuracy**. Don't use cross validation, but instead take at least 25 different **Reasonable** values for **ccp_alpha** and for each of them calculate the **Train** and **Test Accuracies.** Show the corresponding plot of the **cpp_alpha** values versus **Train** and **Test** accuracies. Pick the value of the **ccp_alpha** that correspond to the maximal **Test Accuracy**. Show the optimal value of the parameter with the corresponding **Train** and **Test Accuracies**.

Problem 1.2. score = 10

Train the **Optimal Model** on the entire dataset $\boldsymbol{X_1, y_1}$ (with the optimal value of the **ccp_alpha** parameter). Name it as **Class_A** and show its **Accuracy**. Show the **Classification Report** for the **Class_A** model. For which class it has the best characteristics?

***

Problem 2.1. score = 10

Apply **K-Means** clustering to the entire dataset $\boldsymbol{X_1}$ with **K = 5**. Note that the labels of observations in the clusters, in general, don't correspond to the labels of the same observations in $\boldsymbol{y_1}$ due to some random ordering of the clusters. You need to find the correct correspondence of the labels. For that, for each cluster observations, find their labels from $\boldsymbol{y_1}$ and correct the labels of clusters by the majority vote. You will get a classification model. Name it as **Cluster_A**.

Problem 2.2. score = 10

Comment on the **Cluster_A** model **Accuracy**. Show the **Classification Report** for the **Cluster_A** model. For which class it has the best characteristics? Compare **Class_A** and **Cluster_A** models by the **Accuracies.**

$$X\_2, y\_2 = make\_regression(n\_samples = 5000, n\_features = 20, n\_informative = 3, random\_state = XY, noise = 10)$$

***

Problem 3.1. score = 10

Use **train_test_split** function and split $\boldsymbol{X_2}, \boldsymbol{y_2}$ into 70% - 30% portions, respectively. Apply **SVR** with **kernel = 'rbf'** to the **Train Data** and tune parameters **"gamma"** and **"C"** to maximize the **Score**. Don't use cross validation, but instead take at least 7 different **Reasonable** values for each parameter and calculate the **Train** and **Test Scores**. Pick the values of the parameters that correspond to the maximal **Test Score**. Show the optimal values with the corresponding **Train** and **Test Scores**.

Problem 3.2. score = 10

Train the **Optimal Model** on the entire dataset $\boldsymbol{X_2}, \boldsymbol{y_2}$ (with the optimal values). Name it as **Reg_Full**. Show its **Score**.

***

Problem 3.3. score = 10

Scale $\boldsymbol{X_2}$ features and apply **PCA.** Plot the proportion of explained variance for cumulative components. Take the first **N** principal components that together explain at least **75%** of the original dataset variance. Show **N**. Name, the corresponding dataset as $\boldsymbol{X_2\_PCA\_N}$**.**

Problem 3.4. score = 10

Use **train_test_split** function and split $\boldsymbol{X_2\_PCA\_N}, \boldsymbol{y_2}$ into 70% - 30% portions, respectively. Apply **SVR** with **kernel = 'rbf'** to the **Train Data** and tune parameters **"gamma"** and **"C"** to maximize the **Score**. Don't use cross validation, but instead take at least 6 different **Reasonable** values for each parameter and calculate the **Train** and **Test Scores**. Pick the values of the parameters that correspond to the maximal **Test Score**. Show the optimal values with the corresponding **Train** and **Test Scores**.

Problem 3.5. score = 10

Train the **Optimal Model** on the entire dataset $\boldsymbol{X_2\_PCA\_N}, \boldsymbol{y_2}$ (with the optimal values). Name it as **Reg_PCA**. Show its **Score**.

Problem 3.6. score = 10

Compare models **Reg_Full** and **Reg_PCA** by their **Scores.** Did **PCA** help to increase the **Score** of the model? Are you able to get similar precision with smaller number of features?