Instructions: Please read the following instructions thoroughly

- For the entire assignment, use `Python` for your analysis. Write your code in a Jupyter Notebook named as `[your-student-ID]_hw3.ipynb` (e.g., `2022-20000_hw3.ipynb`). The use of `R` is not allowed. You are allowed to use any libraries in `Python`.
- Type up your report and save as PDF named as `[your-student-ID]_hw3.pdf`. We do not allow the submission of a photo or a scanned copy of hand-written reports.
- Please upload two separate files, your report in PDF and the code in Jupyter Notebook, on eTL **without zipping**. Submissions via email are not allowed. The violation of the filename or submission instruction will result in the penalty of 5 points. (No worries if there are additional numbers at the end of the filename that appear when you submit more than once. That is automatic numbering that the eTL system does on multiple submissions.)
- You can discuss the assignment with your classmates but each student must write up his or her own solution and write their own code. Explicitly mention your classmate(s) you discussed with or reference you used (e.g., website, Github repo) if there is any. If we detect a copied code without reference, it will be treated as a serious violation of the student code of conduct.
- We will apply a grace period of late submissions with a delay of each hour increment being discounted by 5% after the deadline (i.e., 1-minute to 1-hour delay: 95% of the graded score, 1 to 2-hour delay: 90% of the graded score, 2 to 3-hour delay: 85%, so on). Hence, if you submit after 20 hours post-deadline, you will receive 0 points. No excuses for this policy, so please make sure to submit in time.

1. [**25 pts**] We will predict the number of applications received using the other variables in the `College` data set.

   (a) [3 pts] Randomly split the data set into a training set (90%) and a test set (10%). Fit a linear model using least squares on the training set, and report the test error obtained.

   (b) [10 pts] Fit a ridge regression model on the training set, with λ chosen by 10-fold cross-validation using the training set. Report the test error obtained.

   (c) [10 pts] Fit a lasso model on the training set, with λ chosen by 10-fold cross-validation using the training set. Report the test error obtained, along with the number of non-zero coefficient estimates.

   (d) [2 pts] Comment on the results obtained. How accurately can we predict the number of college applications received? Are the models resulting from these three approaches different from each other?

2. [**25 pts**] In this problem, we will use support vector machine (SVM) to predict `default` in `Default` data set attached in the assignment (`Default.csv`). Do not forget to set a random seed before beginning your analysis.

   (a) [10 pts] Fit a SVM that uses `income` and `balance` to predict `default` using the validation set approach, and estimate the test error of this model. In order to do this, you must perform the following steps:

   i Split the sample set into a training set and a validation set.

   ii Fit a SVM using only the training observations.

   iii Obtain a prediction of default status for each individual in the validation set.

   iv Compute the validation set error, which is the fraction of the observations in the validation set that are misclassified.

   (b) [5 pts] Perform 5-fold cross-validation using the same model as in Part (a). Write your own code for the K-fold cross-validation. How does the validation error of the cross-validation the differ from to the results in Part (a)?

   (c) [10 pts] Now consider a SVM that predicts `default` using `income`, `balance`, and a dummy variable for `student`. Estimate the test error for this model using the 5-fold cross-validation set approach. Comment on whether or not including a dummy variable for student would lead to a reduction in the test error rate.

3. [**25 pts**] In this problem, you will perform K-means clustering, with $K = 2$, on a small example with $n = 6$ observations and $p = 2$ features. The observations are as follows.

Observation	X_1	X_2
1	1	4
2	1	3
3	0	4
4	5	1
5	6	2
6	4	0

   (a) [5 pts] Using Jupyter Notebook, plot the observations.

   (b) [5 pts] Randomly assign a cluster label to each observation. Compute the centroid for each cluster. Report the cluster labels for each observation and the centroid for each cluster.

   (c) [5 pts] Assign each observation to the centroid to which it is closest, in terms of Euclidean distance. Report the cluster labels for each observation.

   (d) [5 pts] Repeat computing the centroid for each cluster and (c) until the answers obtained stop changing. In your plot from (a), color the observations according to the cluster labels obtained.

   (e) [5 pts] Repeat the entire sequence from (b) to (d) several times. Do you see the same result as before? Discuss why or why not you observe the same result.

4. [**25 pts**] Consider the `USArrests` data attached in the assignment. You will perform hierarchical clustering on the states.

   (a) [10 pts] Using hierarchical clustering with complete linkage and Euclidean distance, cluster the states using the original data without scaling. Report the result with a dendrogram.

   (b) [5 pts] Cut the dendrogram at a height that results in three distinct clusters. Which states belong to which clusters?

   (c) [5 pts] Hierarchically cluster the states using complete linkage and Euclidean distance, after scaling the variables to have standard deviation one. Report the result with a dendrogram.

   (d) [5 pts] What effect does scaling the variables have on the hierarchical clustering obtained? In your opinion, should the variables be scaled before the inter-observation dissimilarities are computed? Provide a justification for your answer.