CNG 562- Machine Learning

Assignment #1 (10 % of the Actual Mark)

This is a group assignment; max group size is 3.

Given the Iris dataset located in the sklearn library, we need to find the best strategy for finding a linear predictor for the dataset for regression objective.

Your strategy must provide answers for the following aspects:

1- **Data Preprocessing:** Which is better, using the raw data for training or using a scaled data for training (compare between L1 normalization and mean removal preprocessors).

2- **Cross Validation:** Which cross validation type is the best? Compare between random 1-hold out, random 5folds, random 10 fold, and stratified 1-hold out.

3- **Predictor Model:** You need to compare between linear regression and logistic regression predictors.

Hints

- You will need to do different experiments (i.e., 24) to find the best strategy. You can have a separate code file for every experiment or you can use one file, depending on your python skills.

- All of your error comparisons must be as an average of the adopted cross validation type.

- Provide your accuracy results in the following table format for every predictor

	Random 1-Hold out	5 fold	10 fold	Stratified 1-Hold out
Raw Data				
L1 Normalization				
Mean removal				

1) Based on these tables, do we have a winning predictor and a winning strategy? If yes what is it, if no, what shall we do?

2) Based on the winning strategy, provide predication for input [6 3 5 1.5] ? How much is the error?

3) Prepare a small PPT discussing your strategy and findings.

Submit a pdf file that contains all of your winning strategy, your findings/comments, and your code.

Bonus Task [20 marks]

- Draw the ROC curve for your winning strategy comparing linear regression and logistic regression predictors.