Kevin Thomas
CSE 6361-001
1001544593

PROJECT -1 REPORT

Structure of the Code

Packages imported are,

Pandas and Numpy

1) The data is loaded from the external link given . After downloading the data before splitting we are changing the value for the string into integers.
   egs: 'Iris-setosa' → 1 using replace function.
2) Then we shuffle the data so that we can obtain more accuracy while doing cross validation.
3) Then we split the data into Train data and Test data.
4) They are named as A and Y respectively. After that we implement the below formula given from the slide to find the beta,

$$B = ((A\_T*A)**-1)*A\_T*Y$$

5)Then make Predictions in the given model for Linear Regression

6) .Then perform the K fold Cross Validation on the model to overcome the overfitting present in the model .

7) Choose the K value for which the accuracy for the model is high compared to other values.

I have chosen the K value as 5. Its because the accuracy for 5 fold cross validation is more comparatively to a 10 fold or a 3 fold classification.

The accuracy for the 5 fold cross validation has a mean of 96.77 which is higher compared to other folds.