Key Points of the Project

1. **Training a model via linear regression**
   Linear Regression means data is modelled using straight line. Line of Best Fit is predicted.

2. **Using the trained model to do the classification**
   Classification is supervised learning, in which response is categorical. Here, we classify the iris.data based on its three different target (Setosa-0, Virginica-1,Versicolor-2)

3. **Using cross-validation.**
   I use K-fold cross validation here. First, split the dataset into K equal folds. Use fold1 as the testing set and the union of the other folds as training set. Then, calculate testing accuracy.

Code Explanation:

- The code is designed using Python and contains a function for classification problem . First, the input file "iris.data" is read as a dataframe using numpy library.
- Then the three iris species names are changed to '0','1','2' labels in float format and the entire data is manually split into training set and testing set (80/20) .
- With training data as X and training labels as Y we find Beta cap β̂ using linear regression model formula,

$$\hat{\beta} = (A^T A)^{-1} A^T Y$$

- Now, using the estimator values we classify the model and also predict label values over the testing data,

$$f(\mathbf{x}) = w_0 + w_1 x_1 + w_2 x_2 + \ldots w_d x_d = w_0 + \sum_{j=1}^{d} w_j x_j$$

- In order to cross validate we fit the data again into predefined regression model and Compute the predicted values from N-fold cross validation and pass it in regression score function to get the accuracy of predicted output which comes out to be 1.0 mostly.
- Since there is no discrepancy in no. of folds, the N value is chosen to be **10** which is most widely used as the more folds we have, we will be reducing the error due to the bias but increasing the error due to variance which nullifies each other.

RESULT SCREENSHOT