DIKSHA SHARMA
1001679176

PROJECT 1 REPORT
MACHINE LEARNING -6363-001

This project involves the implementation of **multi-variate linear regression** on iris dataset.
We have four independent variables **Petal Length, Petal Width, Sepal Length, Sepal width** and a target/ dependent variable-**Class(Species).We need to predict the class based on these four features using multivariate linear regression.**

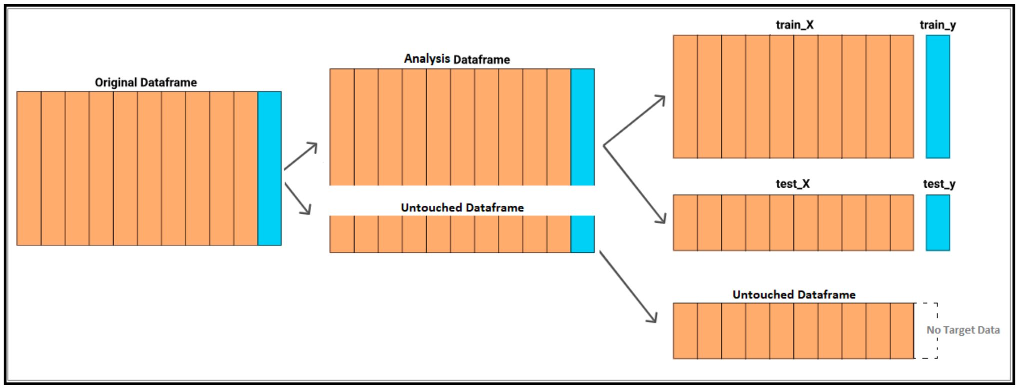
CODE IMPLEMENTATION

Language Used - Python
Environment- Anaconda(Jupyter Notebook & Spyder)

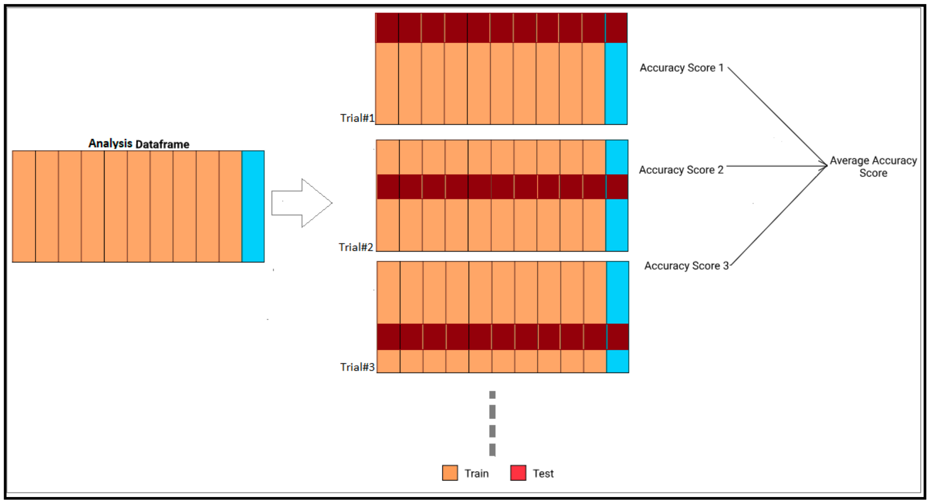
1.Load the dataset with 5 features and 150 rows.
2.Split the data into source(first four features) and target variables(Class).Add a column containing ones to the source data in order to accommodate intercept / constant term in the equation below.

$$Y = a + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k$$

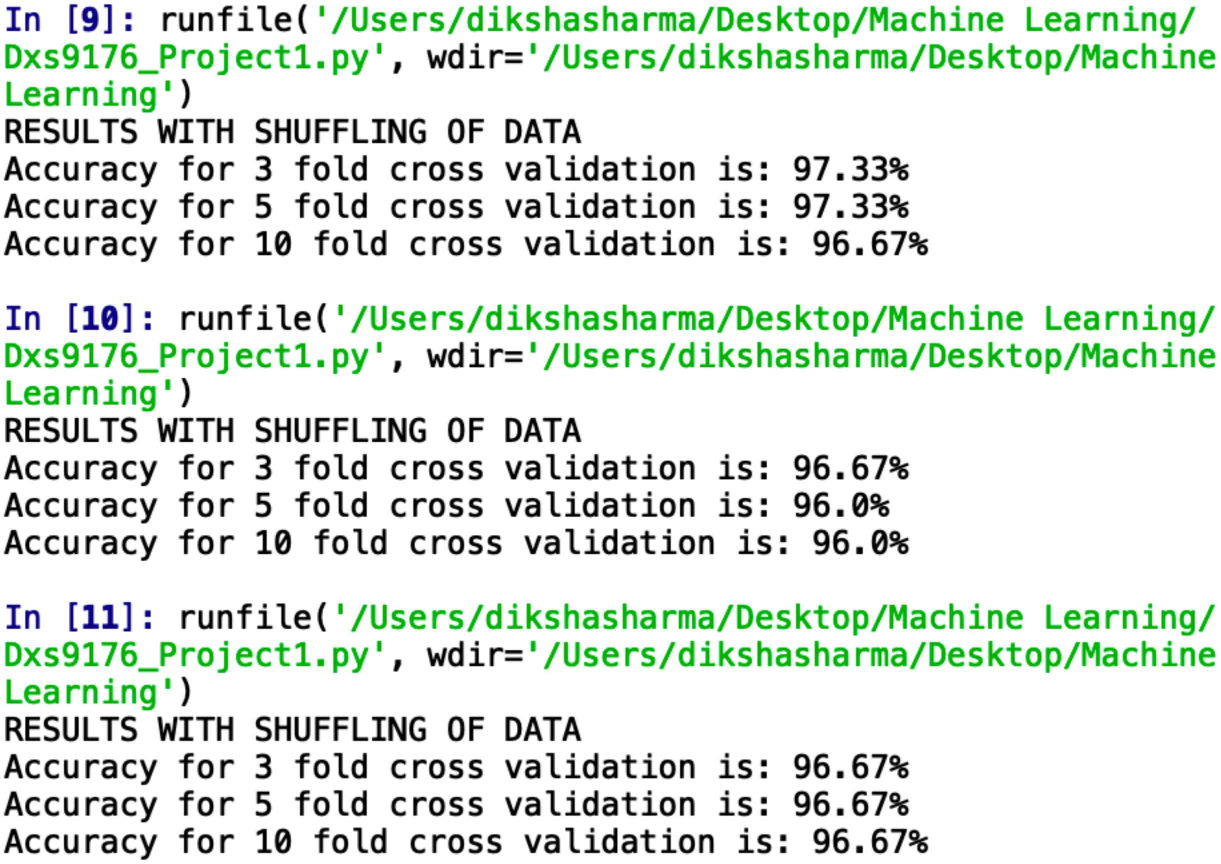
3.Map the target variable strings to integers using dictionary, in order to calculate the beta values by making the same data type for matrix operations.
{'Iris-setosa': 1, 'Iris-versicolor': 2, 'Iris-virginica': 3}
4.Shuffle the data in order to perform k fold cross validation on different values each time the code runs in order to test the accuracy multiple times.



5.Implement k-fold cross validation function with fold values [3,5,10].Used lists and for loops to split the test and the training data for X and y values.Implemented K fold in order to train and test the entire data ,avoiding overfitting.K-Folds cross validation attempts to maximize the use of the available data for training and then testing a model. It is particularly useful for assessing model performance, as it provides a range of accuracy scores across different sets of data.
6.Calculate the Beta values using the formula -> $\boldsymbol{\beta = (X^T X)^{-1} X^T Y}$ and convert the data-frames into matrices for the calculation.Now,the trained model is used for classification.
7.Calculate the accuracy of the obtained Beta values(rounded up to 1 decimal place) on the X_test with the actual target values.
8 .Calculate the mean of the accuracies for each fold cross validation and print the results.



We can see that the best average accuracy for the shuffled data varies each time as there is randomness generated in the data.The best value is for k=3 in the second run as the data is split into 100 samples for training and 50 for testing, providing best results.



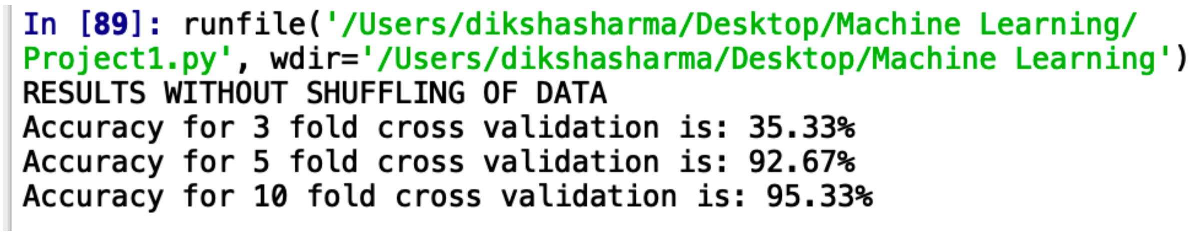
```
In [9]: runfile('/Users/dikshasharma/Desktop/Machine Learning/
Dxs9176_Project1.py', wdir='/Users/dikshasharma/Desktop/Machine
Learning')
RESULTS WITH SHUFFLING OF DATA
Accuracy for 3 fold cross validation is: 97.33%
Accuracy for 5 fold cross validation is: 97.33%
Accuracy for 10 fold cross validation is: 96.67%

In [10]: runfile('/Users/dikshasharma/Desktop/Machine Learning/
Dxs9176_Project1.py', wdir='/Users/dikshasharma/Desktop/Machine
Learning')
RESULTS WITH SHUFFLING OF DATA
Accuracy for 3 fold cross validation is: 96.67%
Accuracy for 5 fold cross validation is: 96.0%
Accuracy for 10 fold cross validation is: 96.0%

In [11]: runfile('/Users/dikshasharma/Desktop/Machine Learning/
Dxs9176_Project1.py', wdir='/Users/dikshasharma/Desktop/Machine
Learning')
RESULTS WITH SHUFFLING OF DATA
Accuracy for 3 fold cross validation is: 96.67%
Accuracy for 5 fold cross validation is: 96.67%
Accuracy for 10 fold cross validation is: 96.67%
```

DIKSHA SHARMA
1001679176

The best accuracy for the data without shuffling is for k=10 as the data is split into 10 sets with 15 samples each and provides the closest approximation with the train data.The training data is 90 % and the testing data is 10 % which is not appropriate for this case.
k=5 is the best value as the data is split into 120 samples for training and 30 samples for testing, which is around 80-20 train test split.



```
In [89]: runfile('/Users/dikshasharma/Desktop/Machine Learning/
Project1.py', wdir='/Users/dikshasharma/Desktop/Machine Learning')
RESULTS WITHOUT SHUFFLING OF DATA
Accuracy for 3 fold cross validation is: 35.33%
Accuracy for 5 fold cross validation is: 92.67%
Accuracy for 10 fold cross validation is: 95.33%
```

REFERENCES

- https://towardsdatascience.com/why-and-how-to-do-cross-validation-for-machine-learning-d5bd7e60c189
- http://mezeylab.cb.bscb.cornell.edu/labmembers/documents/supplement%205%20-%20multiple%20regression.pdf
- https://towardsdatascience.com/data-science-simplified-part-5-multivariate-regression-models-7684b0489015
- https://medium.com/@mtterribile/understanding-cross-validations-purpose-53490faf6a86