Name: Veda Samhitha Manne Student ID: 1002030416

MACHINE LEARNING - CSE 6363-001 - PROJECT 1

Linear Regression:

Linear regression is an analysis which is used to predict the relationship between two variables, i.e., it is used to predict the value of a variable depending on the value of the other variable. In other words, linear relation is a type of supervised learning algorithm which is mostly used in Machine Learning and Data Science.

The main aim of Linear regression is to predict the dependent variable from the values of the independent variables, a linear equation must fit the data. In the real world, many businesses use linear regression to estimate their future cash flow is a common task.

Linear regression is of the form **Y = a + bX,** where X is the explanatory variable and Y is the dependent variable, represents a linear regression line, the slope of the line is b, and a is the intercept.

The Model behind the Linear regression:

- The structural model underlying a linear regression analysis is that the explanatory and outcome variables are linearly related such that the population mean of the outcome for any x value is β0 + β1x.
- The fixed-x, Normality, equal spread, and independent errors assumptions are part of the error model behind a linear regression analysis.
- Repetition of measurements using the same experimental unit is the worst scenario for violating the independent errors assumption in regression.

Simple Linear Regression:

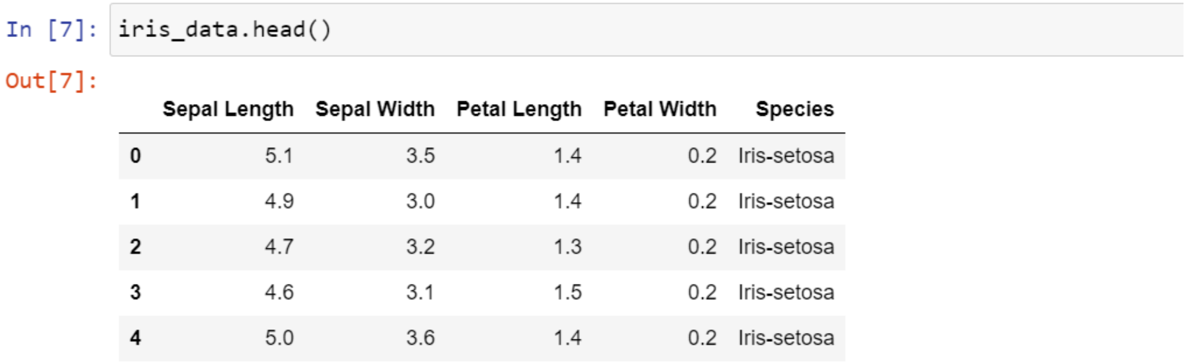
It is a study, that is suitable for a quantitative result and a single quantitative explanatory factor. One can apply simple linear regression, when they want to know the information, In the Simple Linear Regression, the degree of two variables is related.

Linear regression can be used for both simple and multiple regression problems, where the dependent variable is predicted using several independent variables. It is widely used in many different disciplines, including business, engineering, social sciences, and economics, among others.

Sections of Problem:

Simple Linear Regression should be applied with cross validation, to train the model and achieve the classification. We can observe that the data has five columns, and the values of the four feature columns will decide the value of the fifth column, the label column.

There are three sorts of labels column; Iris-setosa, Iris-versicolor, and Iris-virginica.



```
In [7]: iris_data.head()
Out[7]:
```

	Sepal Length	Sepal Width	Petal Length	Petal Width	Species
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa

The above values are measured in centimetres. It is based on the feature columns, should be predicted by our model, and should be able to do so.

Data Pre-processing:

Data processing is the procedure for preparing raw data into the format data and it is the initial and most important stage in building a machine learning model.

The data we utilize is may not be always be well organized, hence it is necessary to structure the data before forming the analysis. For this we have to use the data pre-processing.

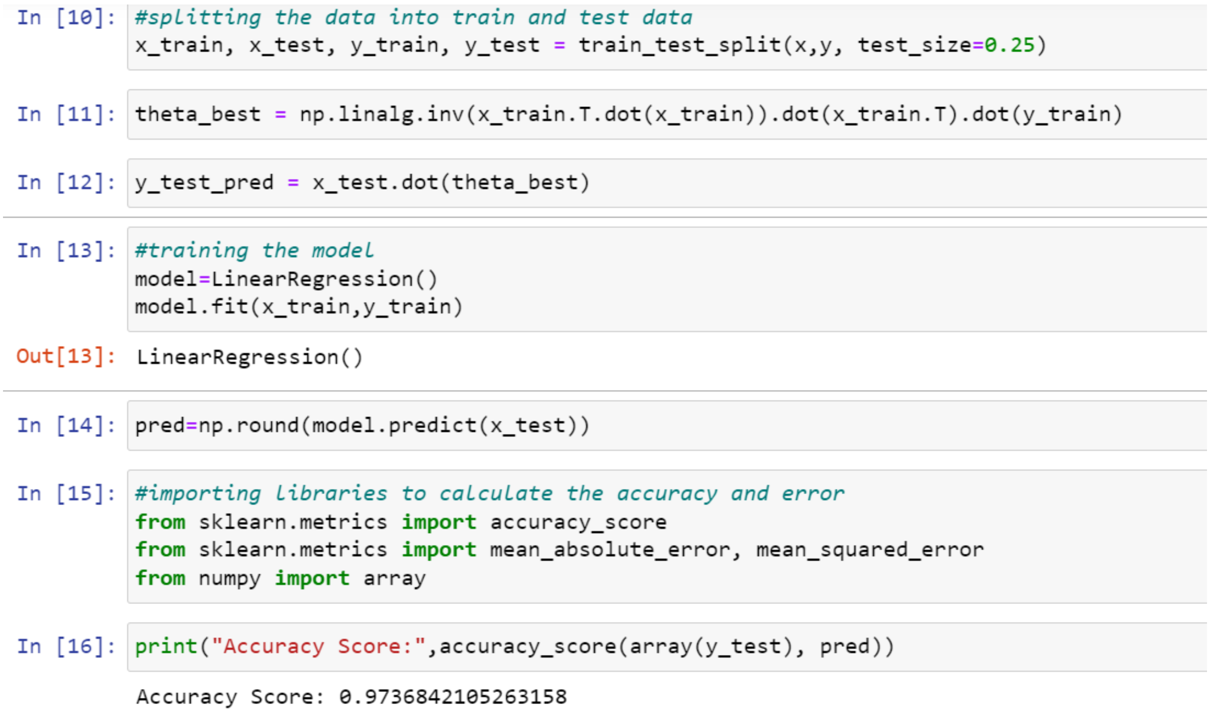
Purpose of data pre-processing:

At certain times, real-world data is not suitable for use in machine learning models because it may not be cleaned, may have missing values, or might have unsuitable format. Cleaning up data and preparing it for machine learning models is a procedure that is essential for increasing the model's efficacy and accuracy.

Training a Model:

It fits the best line to forecast the value of y for a given value of x when training the model. By determining the best 1 and 2 values, the model obtains the best regression fit line. 1 is the number of the intercept. 2 is the x-coefficient. We get the best fit line after we get the best 1 and 2 values.

When the model is being trained, it fits the best line to predict the value of y for a certain value of x. The best regression fit line is obtained by the model by identifying the best 1 and 2 values. The intercept has a value of 1. The x-coefficient equals 2. The best 1 and 2 values are obtained before we obtain the best fit line.



```
In [10]: #splitting the data into train and test data
         x_train, x_test, y_train, y_test = train_test_split(x,y, test_size=0.25)

In [11]: theta_best = np.linalg.inv(x_train.T.dot(x_train)).dot(x_train.T).dot(y_train)

In [12]: y_test_pred = x_test.dot(theta_best)

In [13]: #training the model
         model=LinearRegression()
         model.fit(x_train,y_train)

Out[13]: LinearRegression()

In [14]: pred=np.round(model.predict(x_test))

In [15]: #importing libraries to calculate the accuracy and error
         from sklearn.metrics import accuracy_score
         from sklearn.metrics import mean_absolute_error, mean_squared_error
         from numpy import array

In [16]: print("Accuracy Score:",accuracy_score(array(y_test), pred))

         Accuracy Score: 0.9736842105263158
```

Cross-Validation:

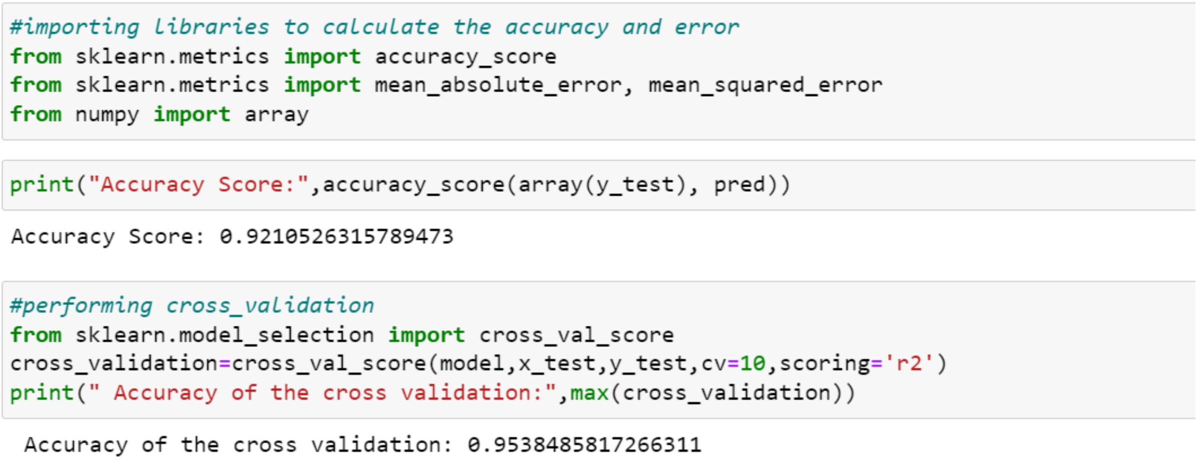
A technique for testing machine learning models called cross-validation entails training several models on subsets of the input data and then assessing them on the complementary subset. Cross-validation can be used to identify overfitting and the inability of a pattern to generalize.

Results:

The Result obtained is attached below.

The accuracy of the Linear Regression Model is: 92.10%

Whereas the accuracy of the Cross Validation is: 95.38%



```
#importing libraries to calculate the accuracy and error
from sklearn.metrics import accuracy_score
from sklearn.metrics import mean_absolute_error, mean_squared_error
from numpy import array
```

```
print("Accuracy Score:",accuracy_score(array(y_test), pred))
```

```
Accuracy Score: 0.9210526315789473
```

```
#performing cross_validation
from sklearn.model_selection import cross_val_score
cross_validation=cross_val_score(model,x_test,y_test,cv=10,scoring='r2')
print(" Accuracy of the cross validation:",max(cross_validation))
```

```
 Accuracy of the cross validation: 0.9538485817266311
```

Reference:

https://www.gs.washington.edu/academics/courses/akey/56008/lecture/lecture9.pdf

https://medium.com/@johnnaujoks/adventures-in-cross-validation-techniques-with-linear-regression-models-75f4e30471

https://www.ibm.com/topics/linear-regression#:~:text=Resources-,What%20is%20linear%20regression%3F,is%20called%20the%20independent%20variable.

https://chat.openai.com/chat