Machine Learning Assignment 2

N.Bhavya

1.a) Linear Regression with One Variables:

Firstly , I took 'all_mcqs_avg_n20' and 'STEP_1' as my data columns . Considering 'STEP_1' is my target variable and 'all_mcqs_avg_n20' is my independent variable. Initially I divided the data set into two parts by splitting the data into 70 % and 30% as my 'Train data ' and 'Test data'. Then by applying Linear regression , regression line is fit on to the data as below.

The optimized coefficient and intercept are calculated by minimizing the cost function with the help of 'Gradient descent algorithm'. Here , I have used a learning rate of 0.001 and no.of.iterations =2000.

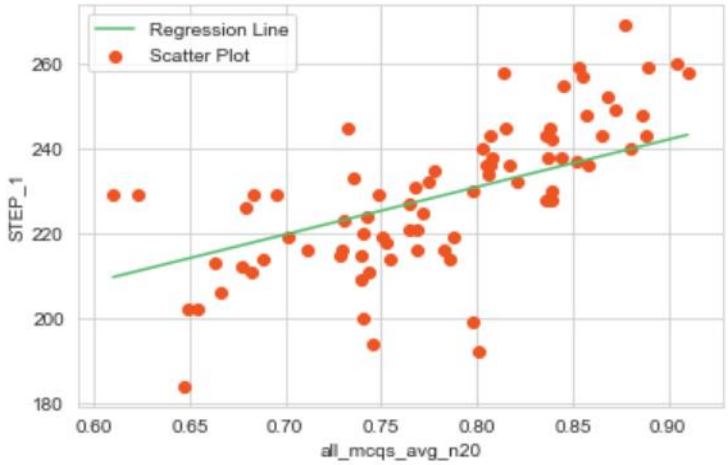
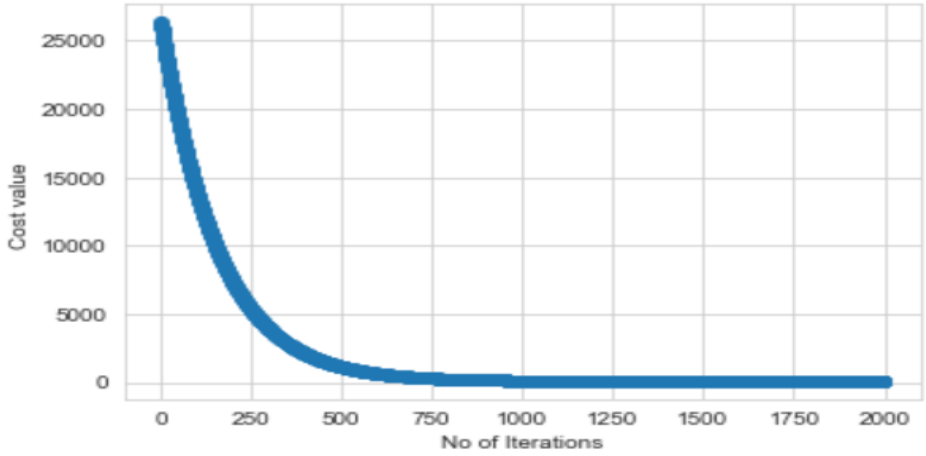


Fig 1. Fitted Regression Line.

The cost function is as below :



From the graph we can observe that the cost value is decreasing from 0 to 2000 iterations.

1B) Performance Of the Model Using Metrics (MSE and R_Square):

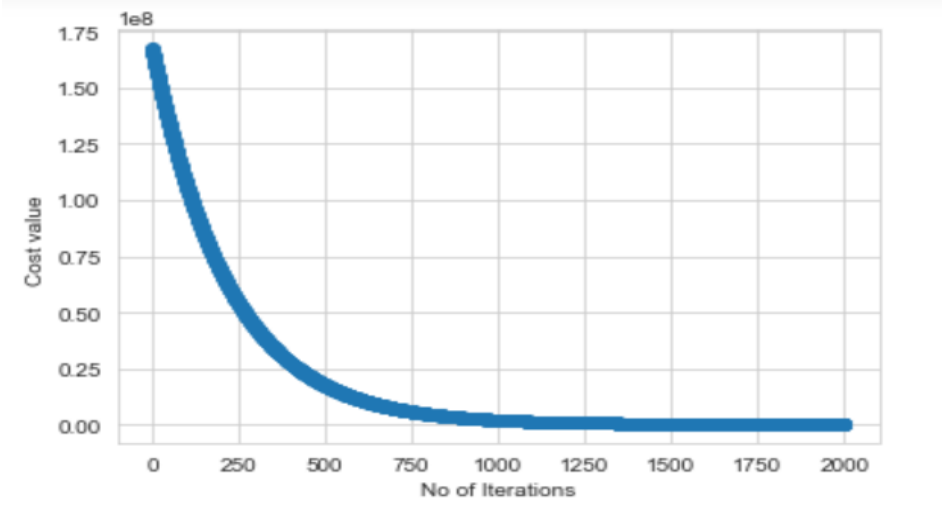
The Mean_square_error , R_Square and Pearson correlation Coefficient value for my model are as below :

```
MSE for single variable  is : 103.33301599088588
r_2value for single variable is  0.4792239778955706
Pearson Correaltion coefficient:  0.7382282395489072
```

2) Linear Regression with Two Variables:

One more column 'all_NBME_avg_n4' is added to the previous data in question 1. Considering 'STEP_1' is my target variable and 'all_mcqs_avg_n20', 'all_NBME_avg_n4' are my independent variables. Initially I divided the data set into two parts by splitting the data into 70 % and 30% as my 'Train data ' and 'Test data'. Then a Linear regression is applied on to the dataset in order to predict the test values .The optimized coefficient and intercept are calculated by minimizing the cost function with the help of 'Gradient descent algorithm'. Here , I have used a learning rate of 0.001 and no.of.iterations =2000.

The cost function is as below :



From the graph we can observe that the cost value is decreasing from 0 to 2000 iterations.

The Mean_square_error ,R_Square and Pearson correlation Coefficient value for my model are as below:

```
MSE for multivariate is : 72.98394570127947
R- square value for multivariate is : 0.6321767195573739
Pearson Correaltion coefficient:  0.8193381413437313
```

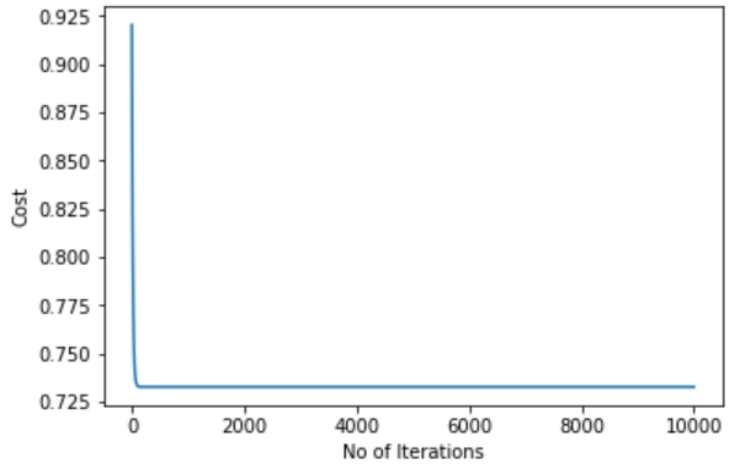
	MSE	R_Square	Pearson Coefficient
Single-Variable	103.33	0.47	0.73
Multiple-Variable	72.98	0.63	0.81

By adding a new column , the performance of the model increased in this case.

3. Logistic Regression with Multiple Variables:

Considering 'LEVEL' is my target variable and 'all_mcqs_avg_n20', 'all_NBME_avg_n4' are my independent variables. Initially I divided the data set into two parts by splitting the data into 70 % and 30% as my 'Train data ' and 'Test data'. Then a Logistic regression is applied on to the dataset in order to predict the class values of unseen data points(Test data points).The optimized coefficients and intercept are calculated by minimizing the cost function with the help of 'Gradient descent algorithm'. Here , I have used a learning rate of 0.1 and no.of.iterations =10000.

Cost function: The cost value decreased from 0 to 10000 iterations



By applying logistic regression, I have obtained an accuracy of 40 % . The confusion matrix, accuracy score, precision , recall and F1 score are mentioned below.

```
The confuision matrix is :
[[ 0  7  0  0]
 [ 0 14  0  0]
 [ 0 12  0  0]
 [ 0  2  0  0]]
The accuracy is  0.4
              precision    recall  f1-score   support

           0       0.00      0.00      0.00         7
           1       0.40      1.00      0.57        14
           2       0.00      0.00      0.00        12
           3       0.00      0.00      0.00         2

   micro avg       0.40      0.40      0.40        35
   macro avg       0.10      0.25      0.14        35
weighted avg       0.16      0.40      0.23        35

f1_score 0.14285714285714288
precision_score 0.1
recall_score 0.25
```

4. Regularization (L2)and Feature Scaling.

A) By applying the feature scaling the performance of the model did not increased here , this is because the given data points are already in similar range i.e.., all the data points are in the range between 0 to 1 **.**

By applying feature scaling with logistic regression, I have obtained an accuracy of 40 % which same as with out applying feature scaling .

The confusion matrix, accuracy score, precision , recall and F1 score are mentioned below.

```
The confuision matrix is :
[[ 0 12  0  1]
 [ 0 14  0  0]
 [ 0  8  0  0]
 [ 0  0  0  0]]
The accuracy is  0.4
              precision    recall  f1-score   support

           0       0.00      0.00      0.00        13
           1       0.41      1.00      0.58        14
           2       0.00      0.00      0.00         8
           3       0.00      0.00      0.00         0

   micro avg       0.40      0.40      0.40        35
   macro avg       0.10      0.25      0.15        35
weighted avg       0.16      0.40      0.23        35

f1_score 0.14583333333333334
precision_score 0.10294117647058823
recall_score 0.25
```

B) I have tested 5 different values of lambda they are, lambda = 0.001,3,6,7.8,9.9

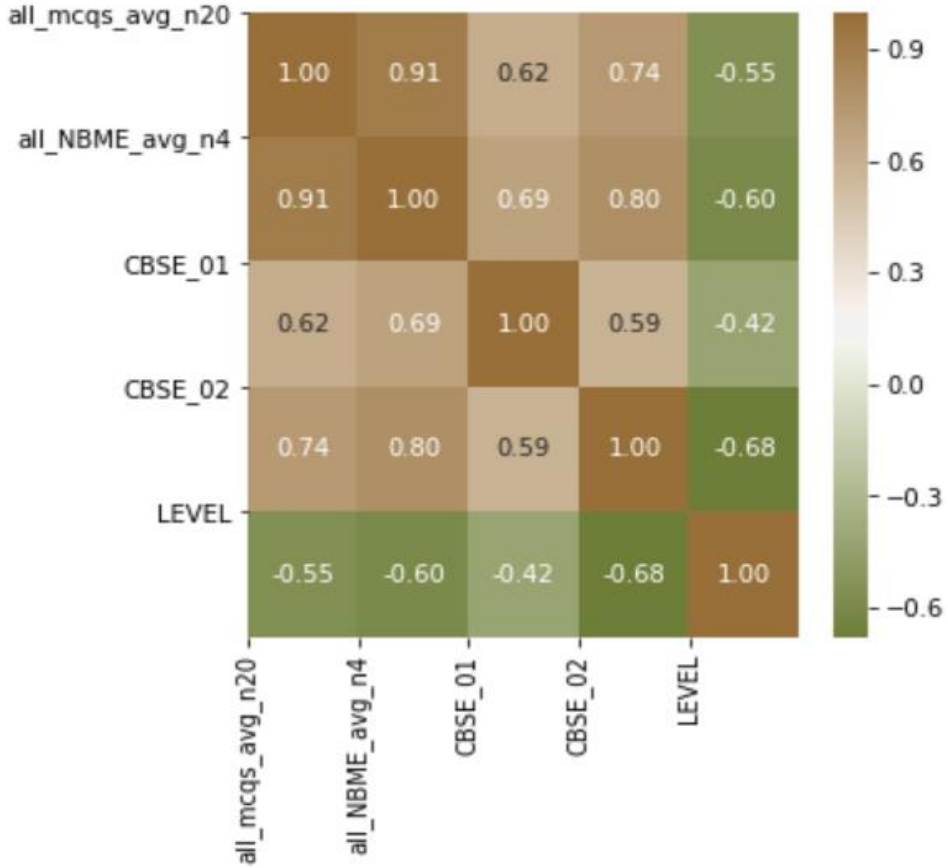
Out of all the best accuracy is obtained for lambda = 0.001.

Therefore with less values of reguralization term the performance of the model is increasing.

```
The accuracy score for lambda = 0.001, alpha =10000,learning rate=0.1  0.45714285714285713
The accuracy score for lambda = 3, alpha =10000,learning rate=0.1 0.42857142857142855
The accuracy score for lambda = 6, alpha =10000,learning rate=0.1 0.4
The accuracy score for lambda = 7.8, alpha =10000,learning rate=0.1 0.4
The accuracy score for lambda = 9.9, alpha =10000,learning rate=0.1 0.4
```

5) Feature Selection :

Correlation matrix:



By using the correlation matrix I choosed three columns since the correlation aong them is less (We have to choose coulms with less correlation value).The columns I choosed are ‘all_mcqs_avg_n20’, ‘CBSE_01’ and ’CBSE_02’ as independent variables and ‘LEVEL’ as my target variable .

Applying feature scaling and regularization techniques:
Here I have applied feature scaling and regularization methods.
I split the data into Train and Test data sets by considering 30 % as my test and 70 % of the data as Train data.
I have used regularization (Lambda) = 0.001(since we obtained 0.001 as best value from question 4).
No.of iterations =10000, Learning_rate = 0.1.

I have obtained an accuracy of 54 % by applying Feature scaling , Feature selection and Regularization .
I have obtained an accuracy of 40 % by with out applying Feature scaling , Feature selection and Regularization .

The accuracy , Confusion matrix, precision, Recall, F1_score obtained by applying Feature scaling , Feature selection and Regularization are as below.

```
For lambda = 0.001, alpha =10000,learning rate=0.1
confusion_matrix
[[6 1 0 0]
 [5 8 1 0]
 [0 7 5 0]
 [0 1 1 0]]
accuracy_score 0.5428571428571428
              precision    recall  f1-score   support

           0       0.55      0.86      0.67         7
           1       0.47      0.57      0.52        14
           2       0.71      0.42      0.53        12
           3       0.00      0.00      0.00         2

   micro avg       0.54      0.54      0.54        35
   macro avg       0.43      0.46      0.43        35
weighted avg       0.54      0.54      0.52        35

f1_score 0.4272778720996038
precision_score 0.4325821237585944
recall_score 0.4613095238095238
```