Q1 a) You are developing a machine-learning model for a prediction task. As you increase the complexity of your model, for example, by adding more features or by including higher-order polynomial terms in a regression model, what is most likely to occur? Explain in terms of bias and variance with suitable graphs as applicable.

- Underfitting : When the model is too simple, it cannot capture the underlying patterns, leading to high bias and low variance. Both training and test errors are high.
- Optimal Complexity : There is a sweet spot where the model complexity is just right. The model captures the underlying patterns without fitting the noise.
- Overfitting : When the model is too complex , it captures the noise in the training data, leading to low bias but high variance. Training error is low, but test error is high.

b)You're working at a tech company that has developed an advanced email

filtering system to ensure users' inboxes are free from spam while safeguarding legitimate messages. After the model has been trained, you are tasked with evaluating its performance on a validation dataset containing a mix of spam and legitimate emails. The results show that the model successfully identified 200 spam emails. However, 50 spam emails managed to slip through, being incorrectly classified as legitimate. Meanwhile, the system correctly recognised most of the legitimate emails, with 730 reaching the users' inboxes as intended. Unfortunately, the filter mistakenly flagged 20 legitimate emails as spam, wrongly diverting them to the spam folder. You are asked to assess the model by calculating an average of its overall classification performance across the different categories of emails.

The performance of the email filtering system was evaluated using these metrics: Precision, Recall, F1-Score, and Accuracy. The results are as follows:

- Precision: 90.9%
- Recall: 80%
- F1-Score: 85.1%
- Accuracy: 93%

These metrics indicate that the system is highly effective at correctly identifying spam emails, with a precision of 90.9%, meaning that the majority of emails flagged as spam are indeed spam.

The recall of 80% shows that the model successfully identifies most spam emails, though we can still improve a bit.

The F1-Score of 85.1% gives a balanced measure of the system's performance, considering both precision and recall.

Finally, the overall accuracy of 93% demonstrates that the system correctly classifies the vast majority of emails.

At last we can say that the email filtering system performs well in maintaining a spam-free inbox.

c)

we know 'm' can be calculated by,

$$m = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$

& 'c' → $$c = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2}$$

for $x = 12$

$\Sigma x = 3 + 6 + 10 + 15 + 18 = 52$

$\Sigma y = 15 + 30 + 55 + 85 + 100 = 285$

$\Sigma xy = 45 + 180 + 550 + 1275 + 1800 \Rightarrow \underline{3660}$

$\Sigma x^2 = 3^2 + 6^2 + 10^2 + 15^2 + 18^2 \Rightarrow \underline{529}$

$$m = \frac{5(3660) - 52(285)}{5(529) - 52^2} \approx 13.7$$

$$c = \frac{529 \cdot 285 - 3660 \cdot 52}{5 \cdot 529 - 52^2} \approx 670.92$$

$$y = (13.65)(12) + (670.92) \rightarrow \cancel{y = 13} \boxed{y \approx 834}$$

d)

f-1 : high degree polynomial regression-model
f-2 : simple regression model (linear)

emperical risk
~~f-1 : lite~~

* f-1 : very low emperical risk & can fit the data very well
* f-2 : it is comparatively less flexible & cannot capture this quadratic data very well.

generalising the model

* although it fits the data perfectly, likely to overfit (f-1)
* as it doesn't work as good as f-1 on training data we can say that it works better on generalised data set (bigger than training).

here the e.g. value of X & Y is taken to be

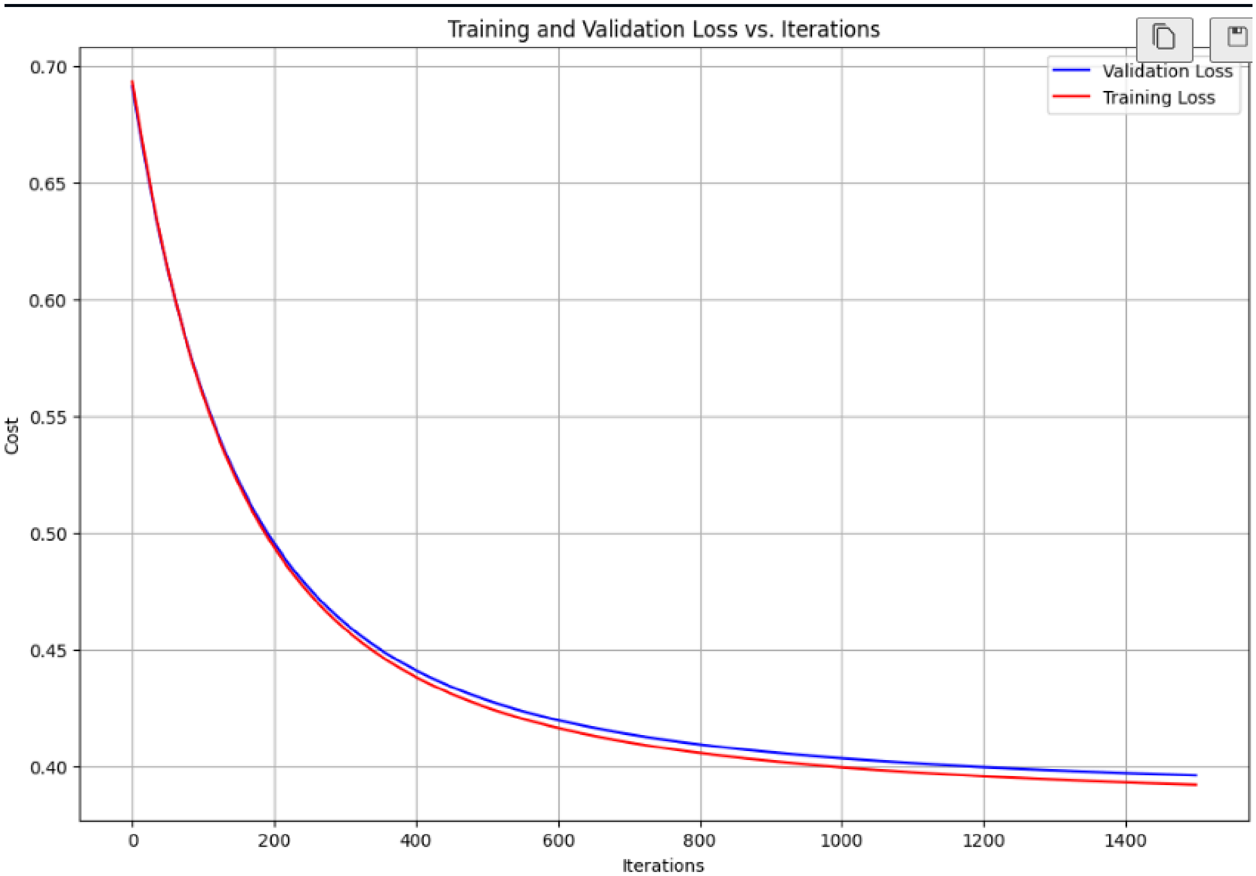
X : {1, 4, 9, 2, 3, 5, 6}
Y : {1, 16, 81, 9, 9, 25, 36}

for testing

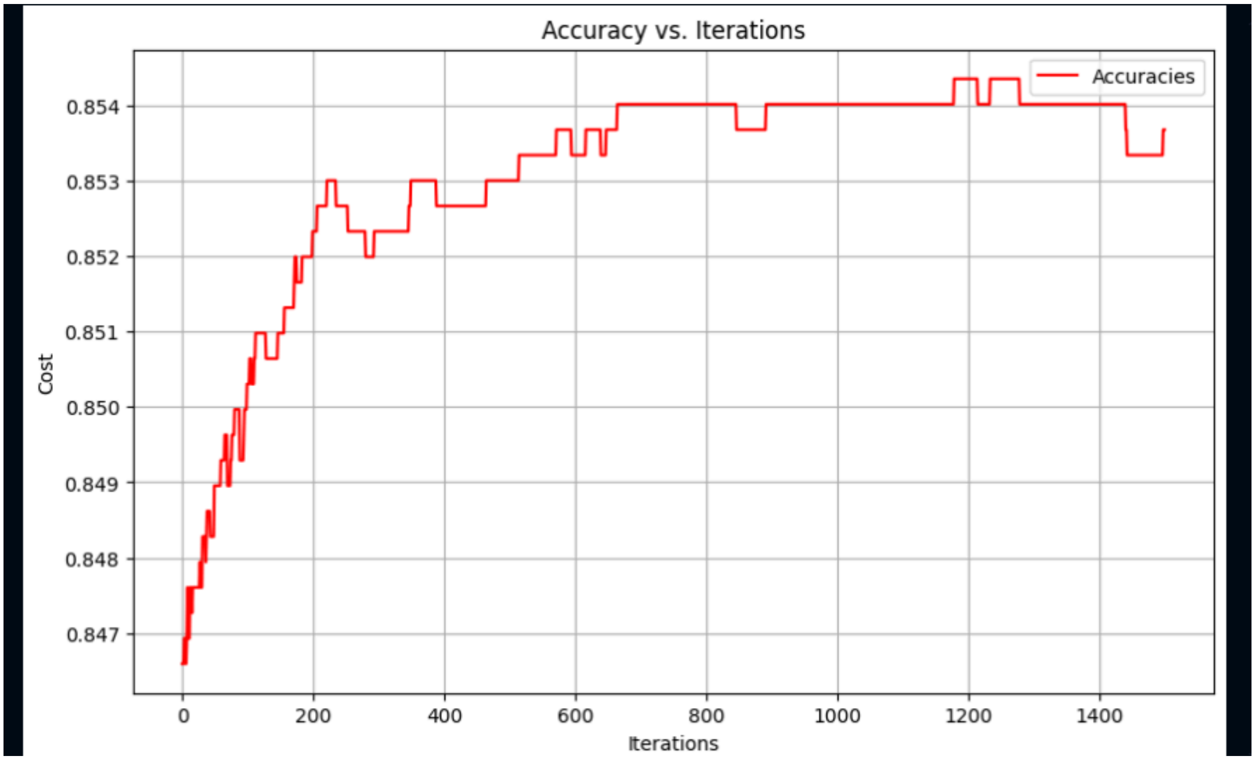
X_test = {4, 6}
Y_test = {1, 1}

f-1 due to overfitting will not predict the correct values.

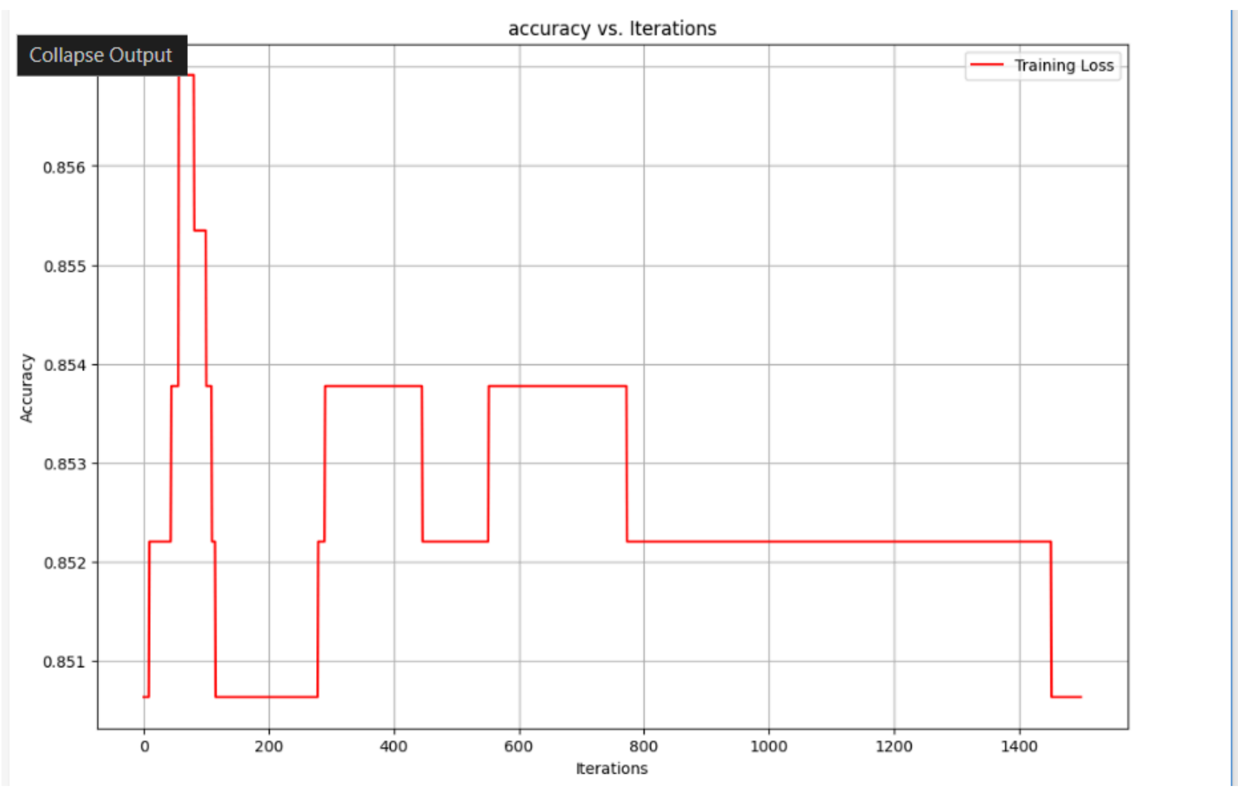
Q2:


Training and Validation Loss vs. Iterations
Validation Loss
Training Loss
Cost
Iterations
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0
200
400
600
800
1000
1200
1400

On training:

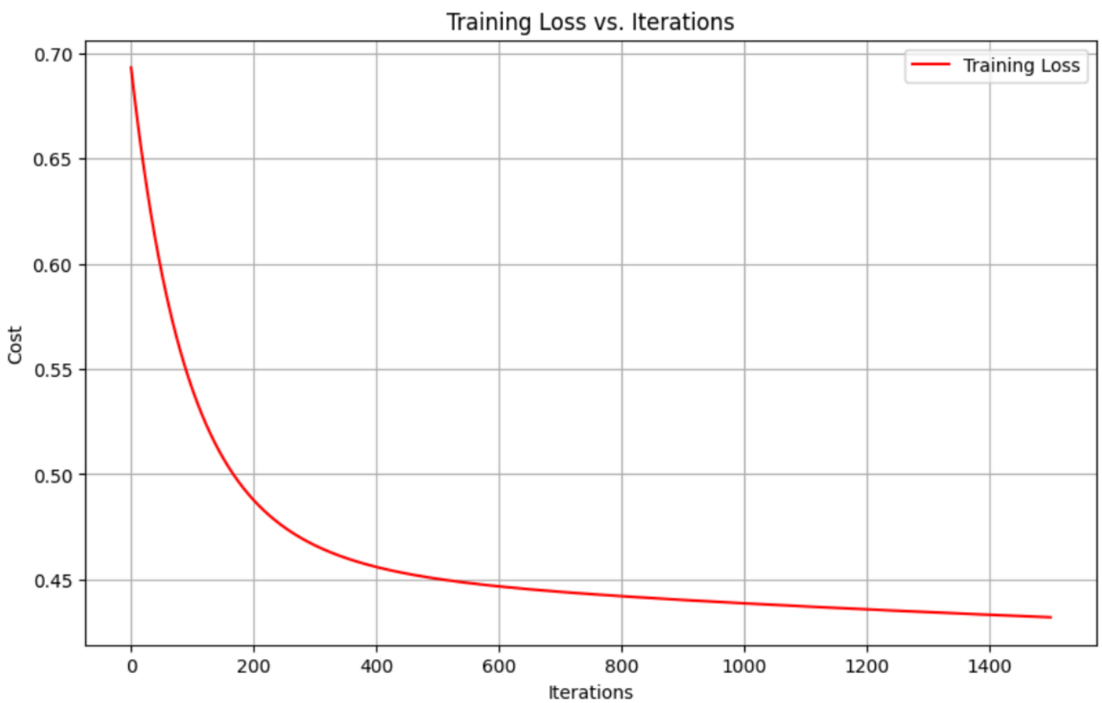


On validation :

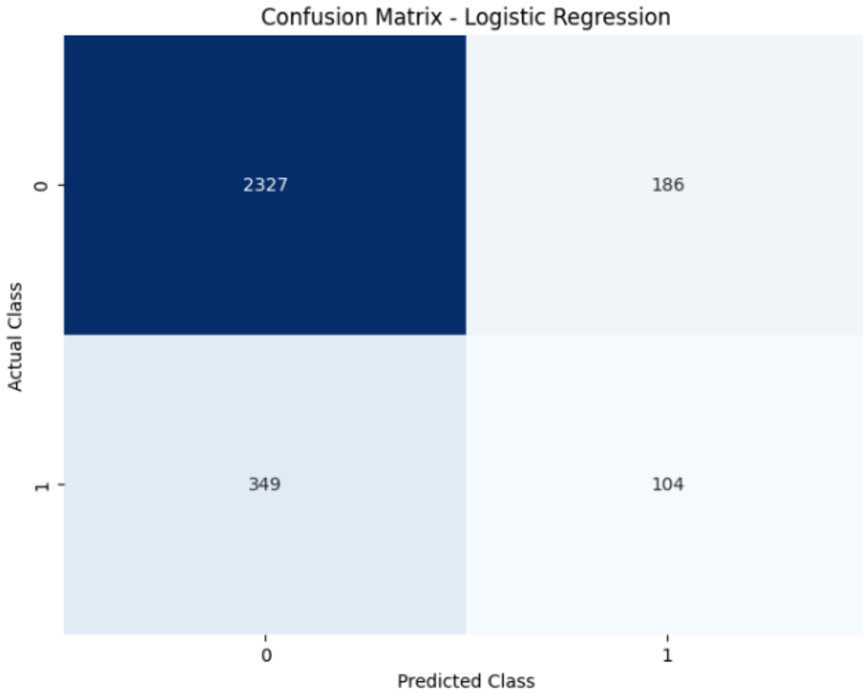


MIN MAX SCALING VS NO SCALING : LOSS VS ITERATION GRAPHS

No scaling accuracy : 0.8569182389937107

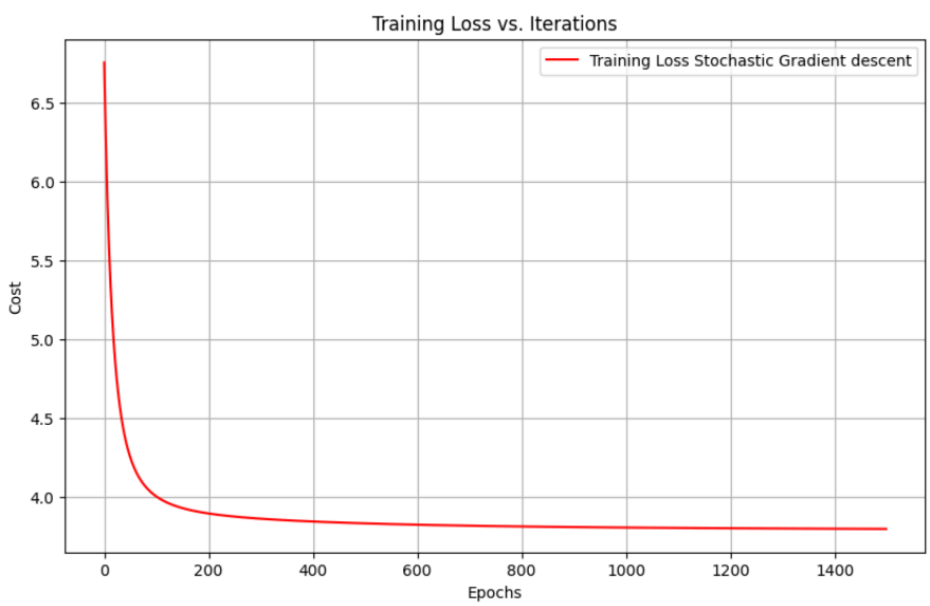


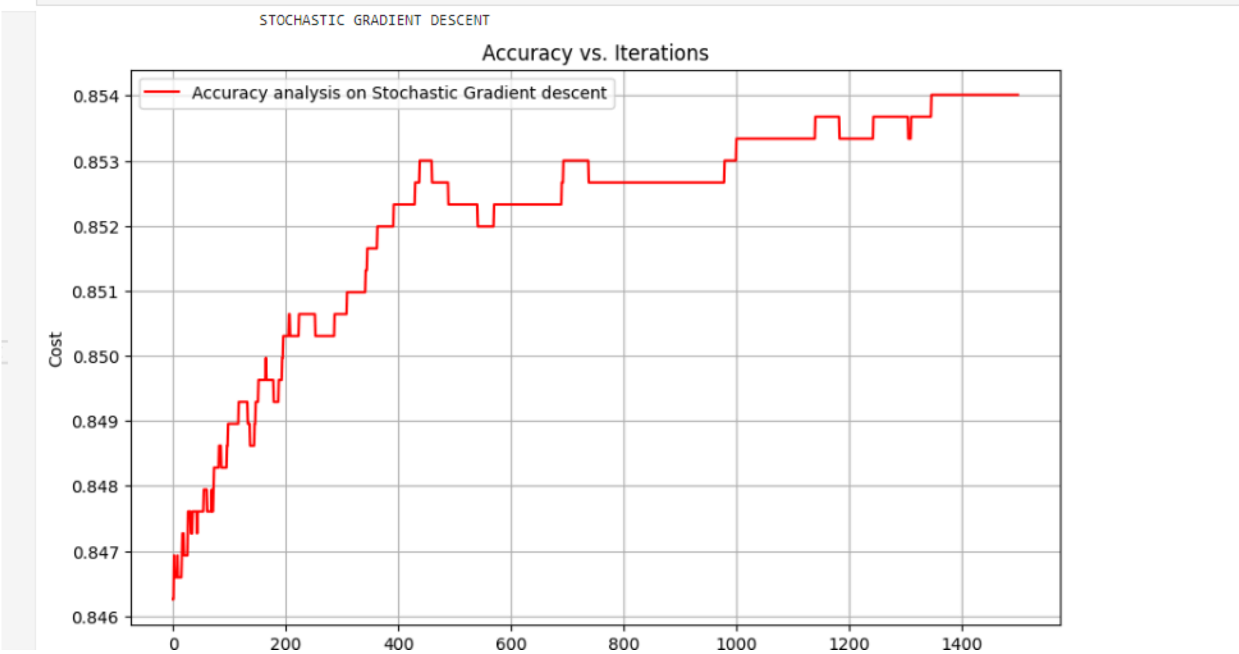
CONFUSION MATRIX



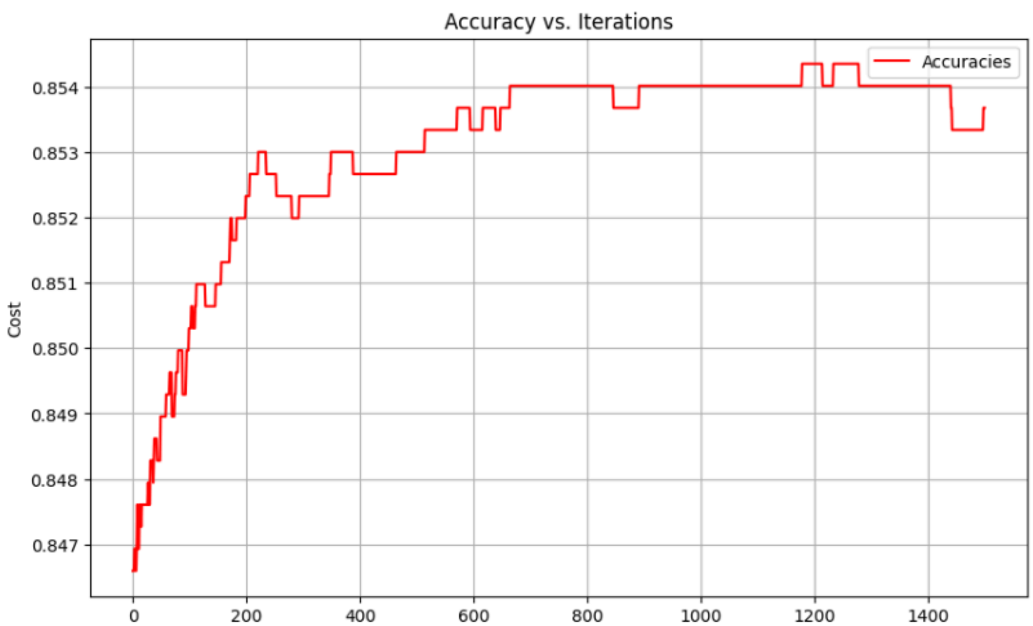

```
 Precision is  : 0.3586206896551724
 Recall score :  0.22958057395143489
F1 score :  0.27994616419919244
ROC-AUC score :  0.5777827262912766
```

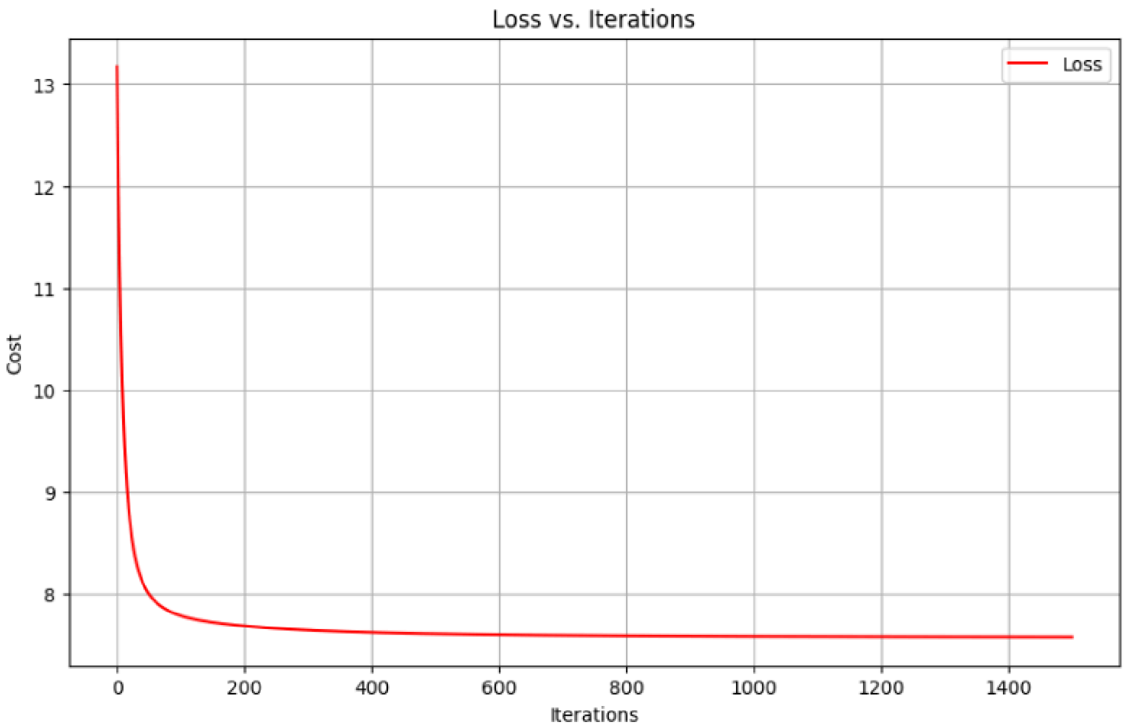
STOCHASTIC GRADIENT DESCENT



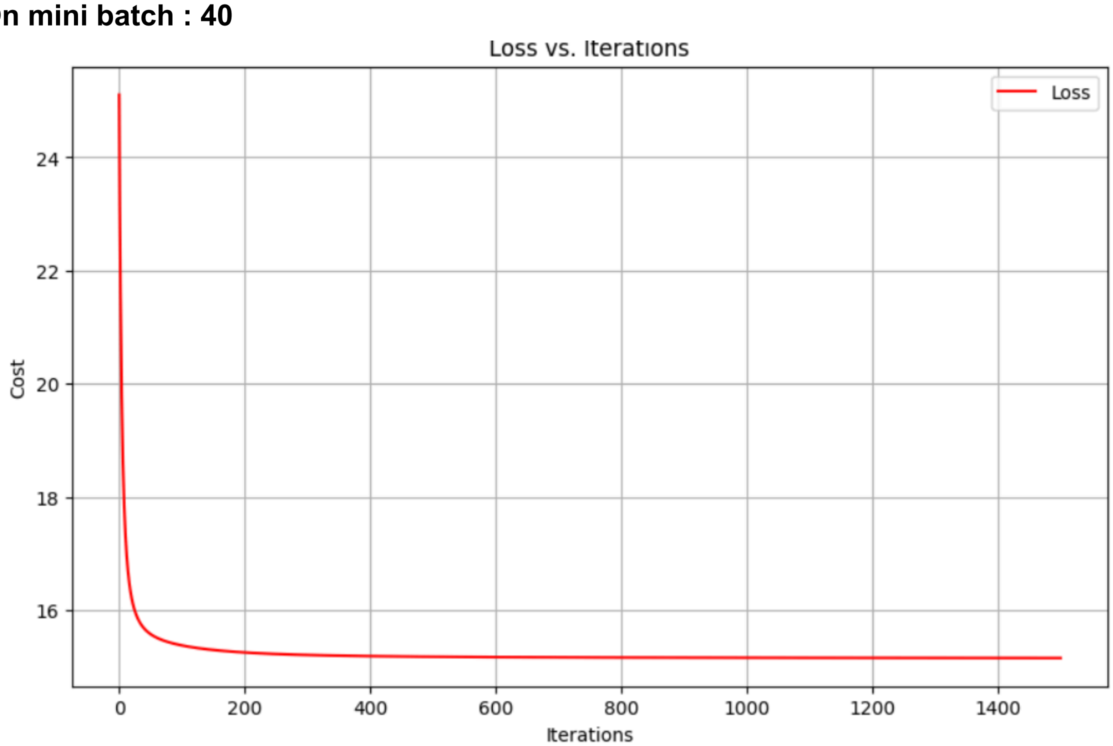


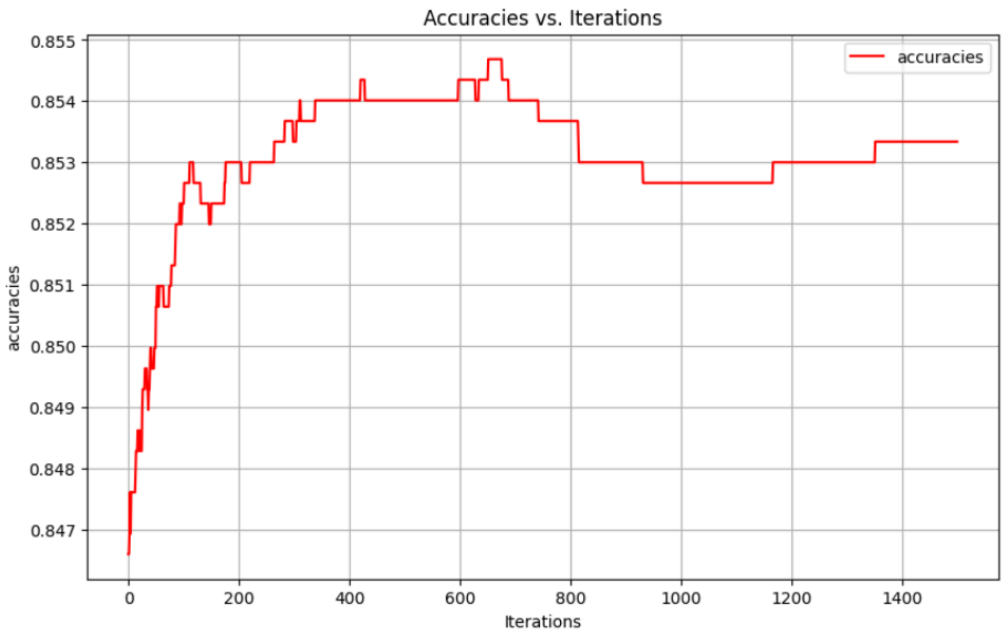
On mini batch :
Batch = 20





On mini batch : 40



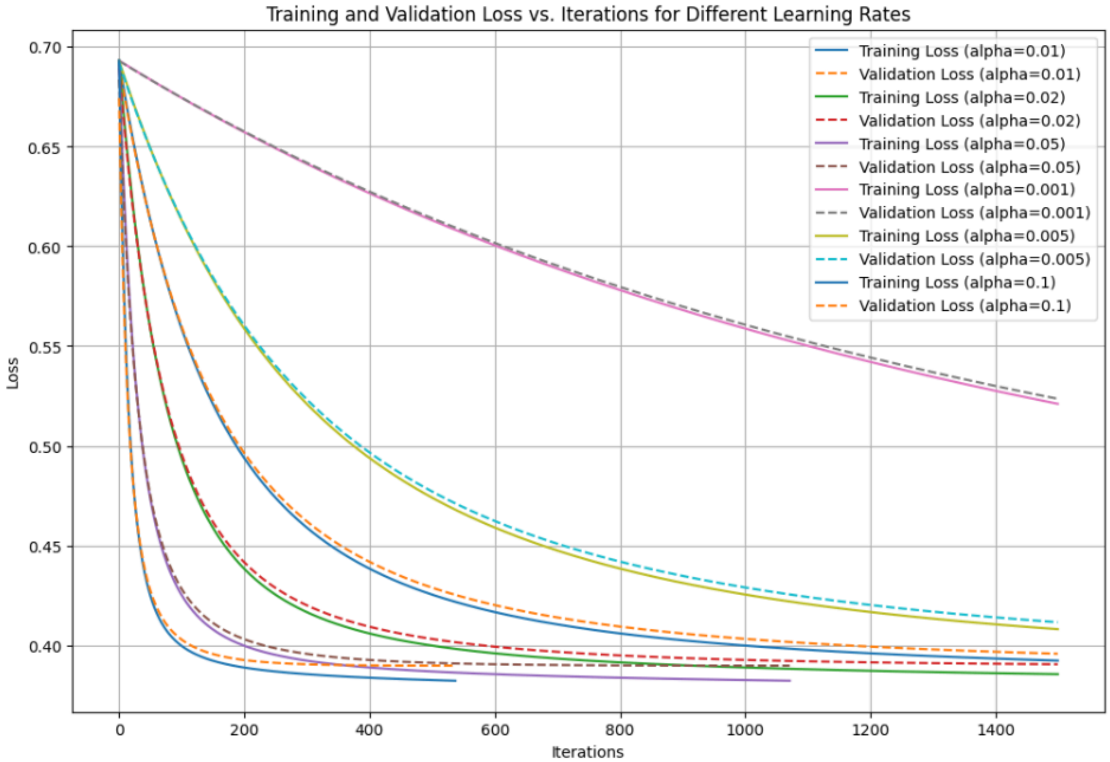


For k =5 , cross validation :

```
                         K-Cross Validation
Mean of accuracy: 0.8462057335581786  standard deviation :   0.013716492011914175
Mean of precisions:  0.5  standard deviation :   0.27386127875258304
Mean of recalls:  0.026739433971546343  standard deviation :   0.008788911870554585
Mean of f1_score:  0.0503816199376947  standard deviation :   0.01674834984685916
```

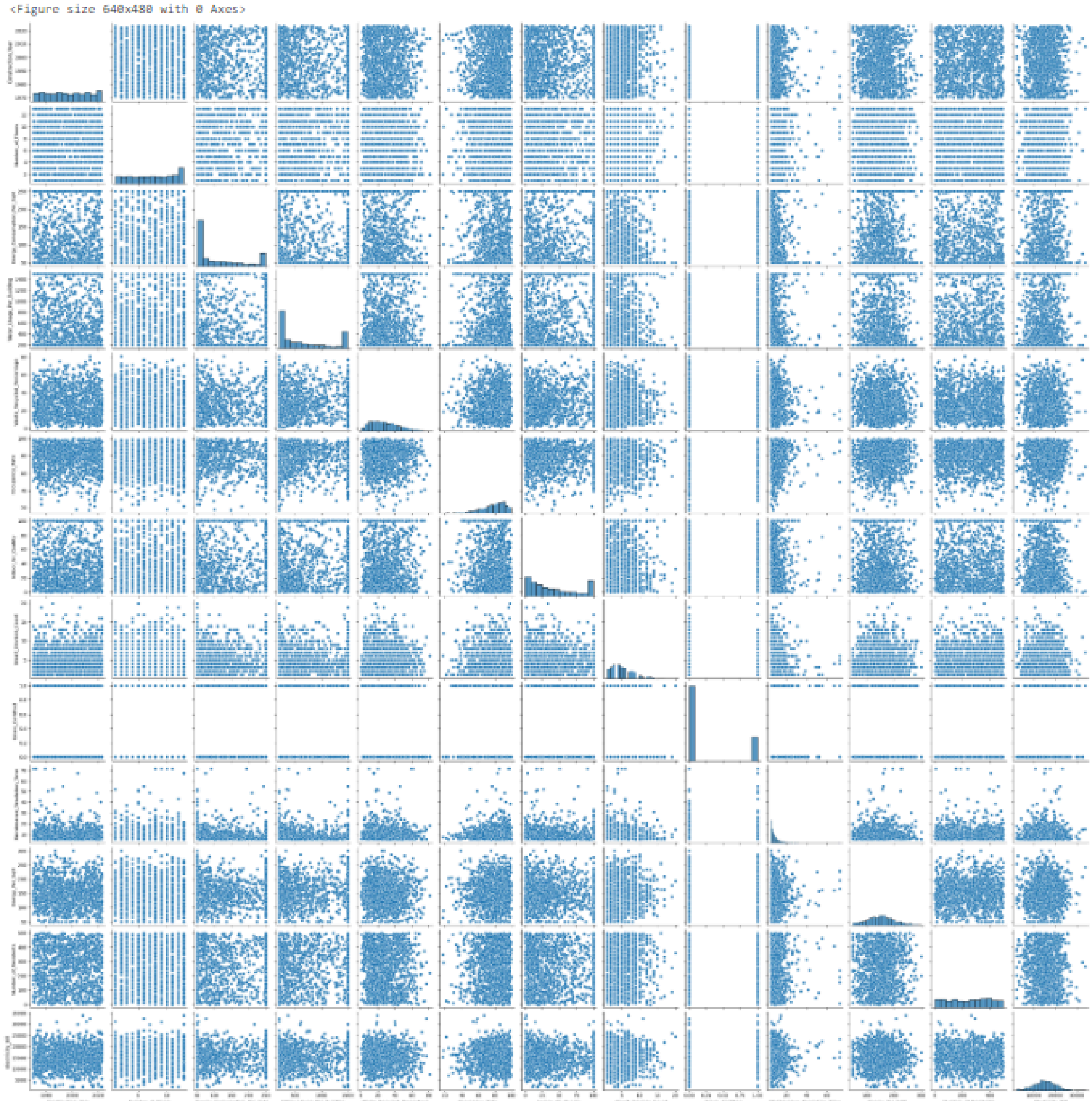
Early stopping with different learning rates :

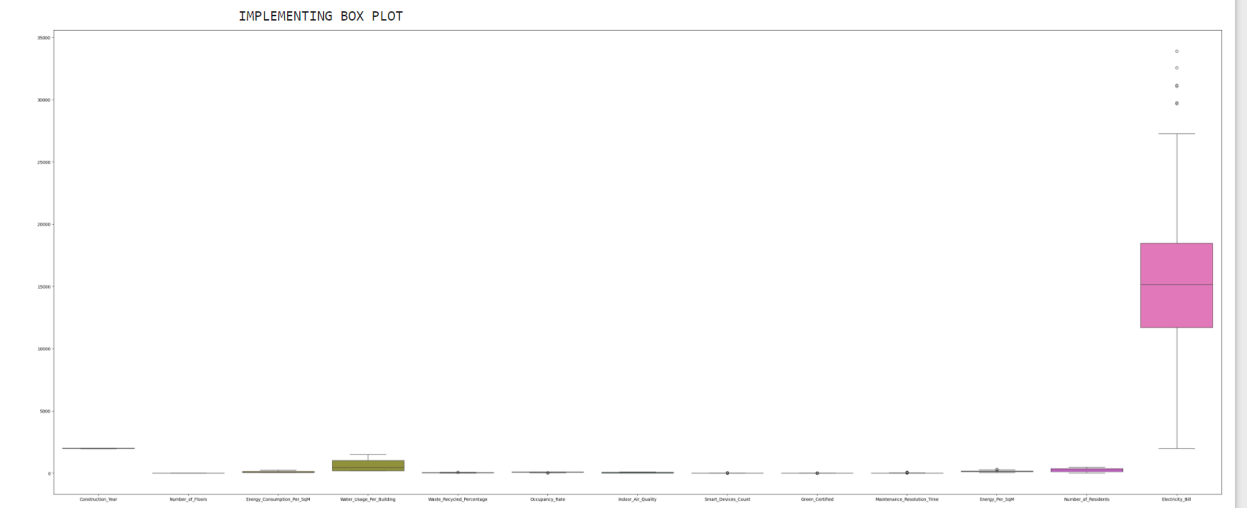
```
IMPLEMENTING EARLY STOP IN GRADIENT DESCENT
Early stop, patience exceeds 5 at iteration 1071
Early stop, patience exceeds 5 at iteration 537
```

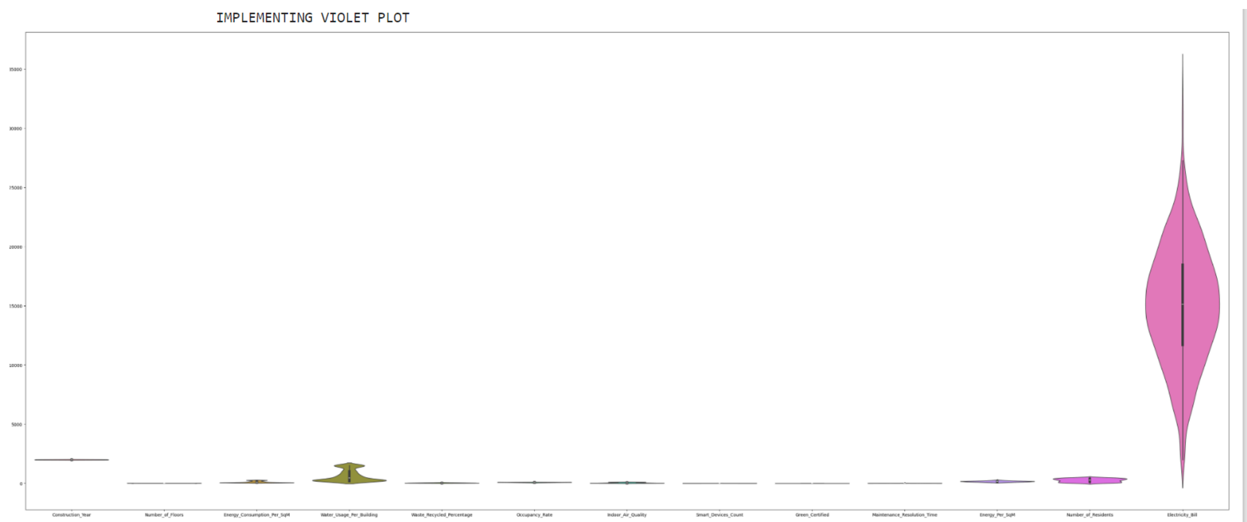


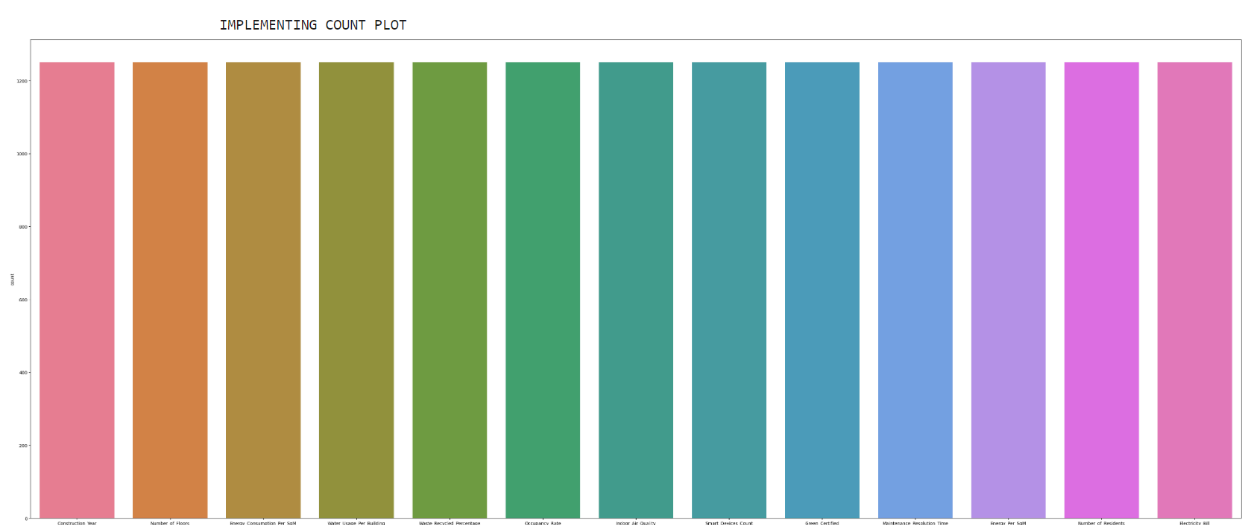
Section C:

```
IMPLEMENTING PAIR PLOT
<Figure size 640x480 with 0 Axes>
```




IMPLEMENTING BOX PLOT
Construction_Year
Number_of_Floors
Energy_Consumption_Per_SqM
Water_Usage_Per_Building
Waste_Recycled_Percentage
Occupancy_Rate
Indoor_Air_Quality
Smart_Devices_Count
Green_Certified
Maintenance_Resolution_Time
Energy_Per_SqM
Number_of_Residents
Electricity_Bill


IMPLEMENTING VIOLET PLOT
Construction_Year
Number_of_Floors
Energy_Consumption_Per_SqM
Water_Usage_Per_Building
Waste_Recycled_Percentage
Occupancy_Rate
Indoor_Air_Quality
Smart_Devices_Count
Green_Certified
Maintenance_Resolution_Time
Energy_Per_SqM
Number_of_Residents
Electricity_Bill


IMPLEMENTING COUNT PLOT
count
Construction_Year
Number_of_Floors
Energy_Consumption_Per_SqM
Water_Usage_Per_Building
Waste_Recycled_Percentage
Occupancy_Rate
Indoor_Air_Quality
Smart_Devices_Count
Green_Certified
Maintenance_Resolution_Time
Energy_Per_SqM
Number_of_Residents
Electricity_Bill

```
                              LINEAR REGRESSION IMPLEMENTATION
Predictions:     [14028.29588778 15485.78388992 14562.77957793 15403.56207339
 16534.02545955]
Labels :         [11586.96964, 7372.100374, 17605.19879, 3160.303673, 16951.74374]
 mean squared error  :  24730978.87656335
 root mean squared error  :  4973.025123258815
 R2 score:  0.024262231392797706
 adjusted R square value :  0.015228426395939021
 mean absolute error :  4013.487414549532
```

Implemented rfse

```
                         IMPLEMENTING RECURSIVE FEATURE ELIMINATION

                         Linear regression on 3 features only
 mean squared error  :  24730978.87656335
 root mean squared error  :  4973.025123258815
 R2 score:  0.024262231392797706
 adjusted R square value :  0.015228426395939021
 mean absolute error :  4013.487414549532
```

```
]: from sklearn.linear_model import Ridge
```

```
                         IMPLEMENTING RIDGE REGRESSION
R2 value is  0.040090229690335044
root mean squared error :  4871.851179963258
mean squared error :  23734933.919709392
mean absolute error  :  3905.5822062999027
adjusted R square value :  0.031353135313531455
```

```
                Implementing independent component analysis ICA


                ACCURACY ANALYSIS OF ICA

        Components  : 5
R2 value is  0.05899205066563529
root mean squared error :  4823.6462825096205
mean squared error :  23267563.45876888
mean absolute error  :  3852.779364638778
adjusted R square value :  0.03135313531353123

        Components  : 4
R2 value is  0.031969988530254345
root mean squared error :  4892.414203263112
mean squared error :  23935716.736290626
mean absolute error  :  3914.8369053263605
adjusted R square value :  0.03135313531353123

        Components  : 6
R2 value is  0.08862536834413115
root mean squared error :  4747.087919892537
mean squared error :  22534843.71918965
mean absolute error  :  3791.596891200691
adjusted R square value :  0.03135313531353123

        Components  : 8
R2 value is  0.27855794313860704
root mean squared error :  4223.568434931977
mean squared error :  17838530.324553754
mean absolute error  :  3405.2455641533816
adjusted R square value :  0.03135313531353123
```

```
print("Gradient Boosting Regressor Metrics:", gbr_metrics)
```

```
                        ELASTICNET REGULARIZATION

ElasticNet with alpha=0.1
ElasticNet Coefficients:  [-188.05334957 -194.7530676   250.0206893   149.62509516  -84.20088495
   84.56093749 -164.31000627 1026.71689113  163.71320324 -194.69104001
  220.5003022  -199.86776472   39.37244414  160.17034823  101.81470526]
ElasticNet Intercept:  14758.796183976001
Evaluation on test set:
MSE: 20508488.89608697
RMSE: 4528.629913791474
MAE: 3618.572646802898
R²: 0.07246248211724537
Adjusted R²: 0.013004948919632997

ElasticNet with alpha=0.5
ElasticNet Coefficients:  [-132.55342793 -155.97143848  192.66355021  125.4147739   -27.64467219
   37.85129546 -121.54855902  750.90073435  109.26283983 -142.66845703
  164.51953988 -138.57865253   23.55073421  129.45768849   59.79540689]
ElasticNet Intercept:  14758.796183976001
Evaluation on test set:
MSE: 20669973.915639374
RMSE: 4546.424300000977
MAE: 3629.495658462251
R²: 0.06515899842472106
Adjusted R²: 0.005233293195536559

ElasticNet with alpha=1.0
ElasticNet Coefficients:  [ -97.06805171 -123.5500937   150.31428296  102.87676612   -2.82325032
   16.58464225  -91.62093686  564.79056227   77.15229828 -106.4330244
  125.71581541  -99.76328586   13.84536472  103.97680504   36.75932753]
ElasticNet Intercept:  14758.796183976001
Evaluation on test set:
MSE: 20884815.71133026
RMSE: 4569.990778035582
MAE: 3646.892010577539
R²: 0.055442347582121854
Adjusted R²: -0.005106219880562701

ElasticNet with alpha=5.0
ElasticNet Coefficients:  [-31.02134617 -45.43071145  54.87150212  41.09037866  11.57462869
  -1.93402607 -30.74166746 191.60246902  23.07165486 -34.56602991
  44.35256501 -30.26655203   0.95897541  40.01359262   6.23290908]
ElasticNet Intercept:  14758.796183976001
Evaluation on test set:
MSE: 21591219.97389798
RMSE: 4646.635339027367
MAE: 3708.5851760396854
R²: 0.023493798879010153
Adjusted R²: -0.03910275247489947
```