SUYASH KUMAR
2021293

ML Assignment-1
REPORT

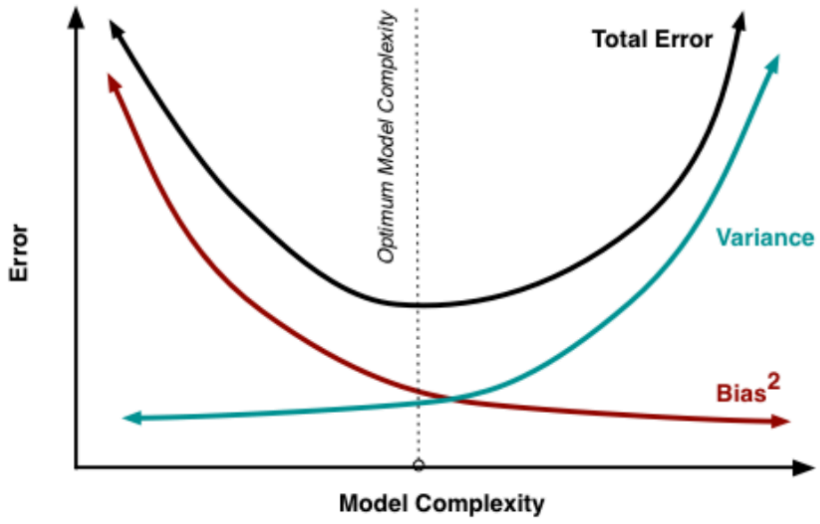
Ques-1| Part - a

As the complexity of a machine-learning model increases (e.g., by adding more features or using higher-order polynomial terms), the model keeping learning every single detail of the training dataset and it get prone to **Overfitting**

1. Low Complexity Model: **High Bias** (**Underfitting**) as the model oversimplifies the data, leading to both high training and test error. **Variance** is low since the model is stable and not sensitive to fluctuations in the data.

2. Moderate Complexity Model: **Balanced Bias and Variance,** where the model captures the underlying patterns without overfitting or underfitting. This leads to the lowest generalization error.

3. High Complexity Model: **Low Bias** (**Overfitting**), as the model fits the training data very well, but **High Variance** as it becomes sensitive to noise, resulting in poor generalization to new data.



In a **bias-variance tradeoff graph**, bias decreases and variance increases with model complexity. The total error follows a U-shape, with the optimal point at balanced complexity for minimal generalization error.

Ques -1| Part - b

To assess the performance of the email filtering model, we calculate key metrics based on its classification outcomes:

True Positives (TP): 200 spam emails correctly identified.
False Negatives (FN): 50 spam emails misclassified as legitimate.
True Negatives (TN): 730 legitimate emails correctly identified.
False Positives (FP): 20 legitimate emails wrongly flagged as spam.

Key metrics:
1. Precision (accuracy of spam detection):

Precision = (TP)/ (TP + FP) = 200/(200 + 20) = 0.909

2. Recall (ability to detect all spam):

Recall = TP/ (TP + FN) = 200/ (200 + 50) = 0.8

3. Accuracy (overall correct classifications):

Accuracy = (TP + TN)/(TP + TN + FP + FN) = (200 + 730)/(200 + 730 + 20 + 50) = 0.93

4. **F1 Score** = 2 x (Precision x Recall)/ (Precision + Recall)
= 2(0.909 x 0.8)/ (0.909 + 0.8)
= 0.85

The model exhibits strong performance with high precision, reasonable recall, and an F1 score of 0.85, indicating a good balance between identifying spam and minimizing false positives.

Ques-1| Part - c

Page No.

Ques-1) part - (c)

We have to equation for simple linear regression on our dataset

Eqn for simple linear regression:

$$y = mx + c$$

So, we have to calculate variables m & c

where m is the y-intercept & c is slope.

$$m = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$\bar{x}$ → mean of x values

$\bar{y}$ → mean of y values

$$\bar{x} = \frac{3+6+10+15+18}{5} = \frac{52}{5} = 10.4$$

$$\bar{y} = \frac{15+30+55+85+100}{5} = \frac{285}{5} = 57$$

$$m = \frac{(3-10.4)(15-57)}{(3-10.4)^2} + \frac{(6-10.4)(30-57)}{(6-10.4)^2}$$

$$+ \frac{(10-10.4)(55-57)}{(10-10.4)^2} + \frac{(15-10.4)(85-57)}{(15-10.4)^2}$$

$$+ \frac{(18-10.4)(100-57)}{(18-10.4)^2}$$

$m = 5.78$

$c = \bar{y} - m\bar{x} = 57 - 5.78 \times 10.4$

$= -3.11$

So equation for linear regression

$y = 5.78x - 3.11$

at $x = 12$

$y = 5.78(12) - 3.11 = \boxed{66.25}$

Ques-1| Part - d

Consider a toy example with a dataset of 10 points:

Training Data (X, Y):
X = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
Y = [2, 4, 6, 8, 10, 12, 14, 16, 18, 20]

Model f1 is a 9th-degree polynomial, which perfectly fits all points in the training data (zero training error, i.e., very low empirical risk). However, it overfits, capturing noise in the data.

Model f2 is a simple linear model:

$$Y = 2X$$

It does not perfectly fit the data (higher training error), but it captures the true underlying relationship.

In this case, **f1** has lower empirical risk on the training set but will likely perform poorly on new data due to overfitting. **f2**, with a higher empirical risk on the training set, may generalize better to unseen data because it captures the underlying trend without overfitting.

Ques -2

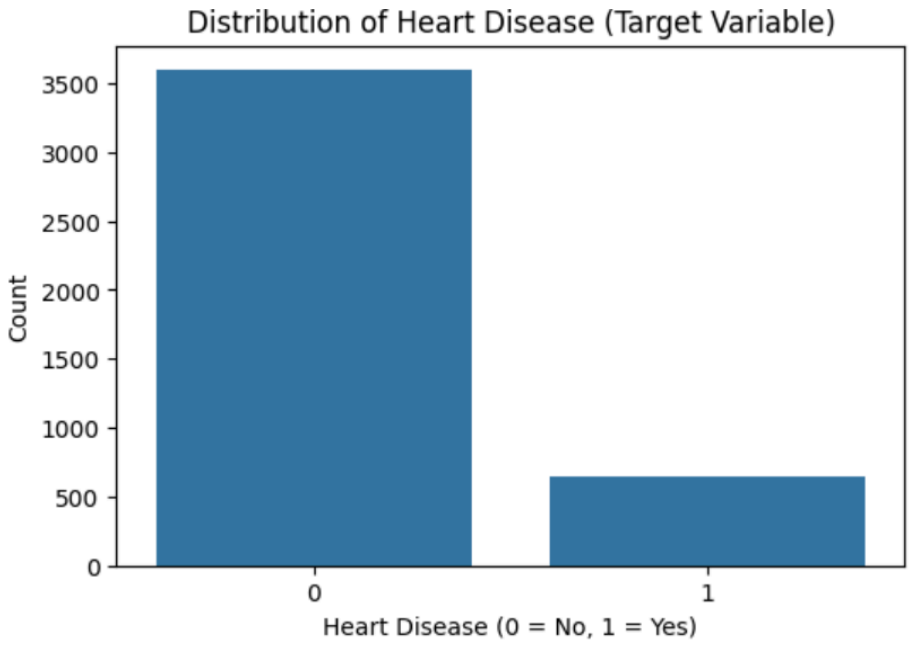
EDA on the Dataset:

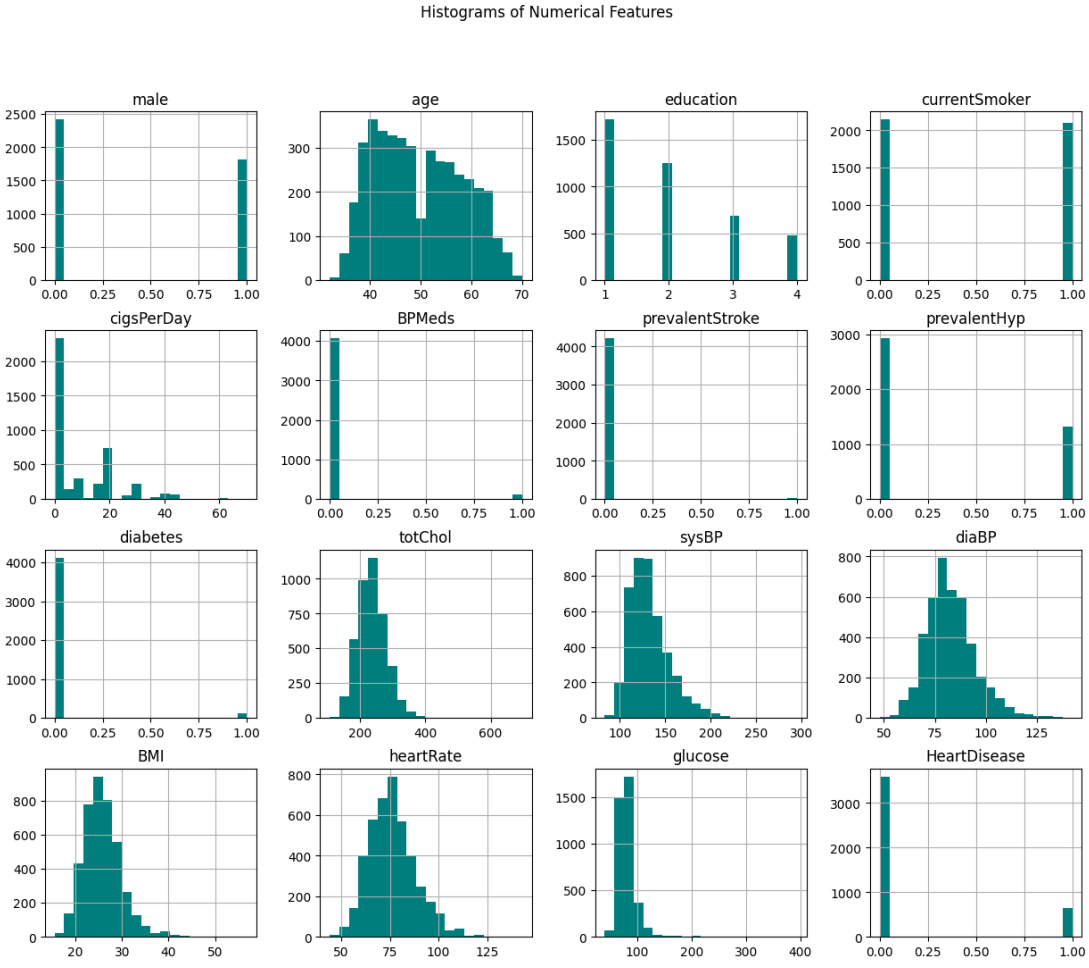
```
First 5 rows of the dataset:
   male  age  education  currentSmoker  cigsPerDay  BPMeds  prevalentStroke  \
0     1   39        4.0              0         0.0     0.0                0
1     0   46        2.0              0         0.0     0.0                0
2     1   48        1.0              1        20.0     0.0                0
3     0   61        3.0              1        30.0     0.0                0
4     0   46        3.0              1        23.0     0.0                0

   prevalentHyp  diabetes  totChol  sysBP  diaBP    BMI  heartRate  glucose  \
0             0         0    195.0  106.0   70.0  26.97       80.0     77.0
1             0         0    250.0  121.0   81.0  28.73       95.0     76.0
2             0         0    245.0  127.5   80.0  25.34       75.0     70.0
3             1         0    225.0  150.0   95.0  28.58       65.0    103.0
4             0         0    285.0  130.0   84.0  23.10       85.0     85.0

   HeartDisease
0             0
1             0
2             0
3             1
4             0
```

Null values in the dataset:

```
Missing Values in Each Column:
male                0
age                 0
education         105
currentSmoker       0
cigsPerDay         29
BPMeds             53
prevalentStroke     0
prevalentHyp        0
diabetes            0
totChol            50
sysBP               0
diaBP               0
BMI                19
heartRate           1
glucose           388
HeartDisease        0
dtype: int64
```

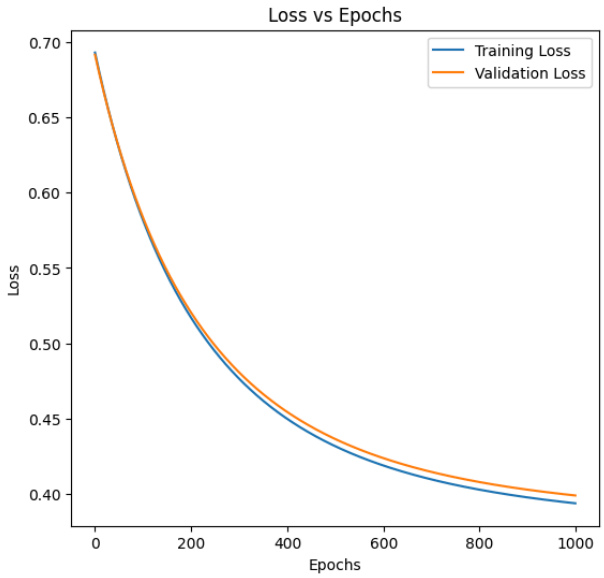
Replaced the Null values with the median of each Row.

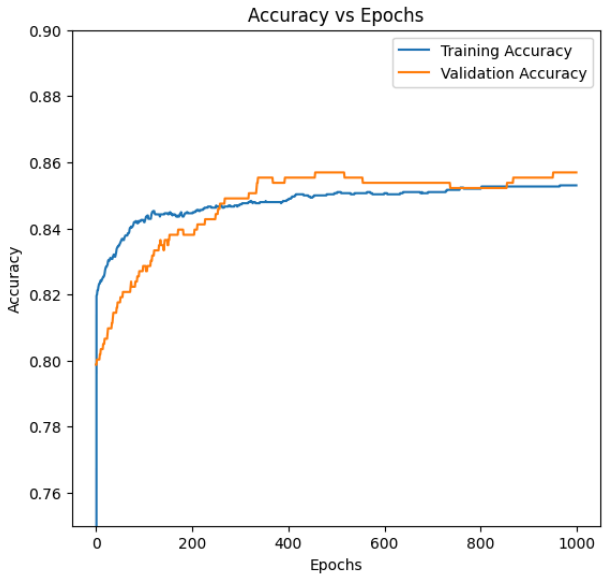

Distribution of Heart Disease (Target Variable)
Count
Heart Disease (0 = No, 1 = Yes)


Histograms of Numerical Features
male
age
education
currentSmoker
cigsPerDay
BPMeds
prevalentStroke
prevalentHyp
diabetes
totChol
sysBP
diaBP
BMI
heartRate
glucose
HeartDisease

Ques-2 | Part - a

```
Epoch 0: Train Loss = 0.6931, Validation Loss = 0.6917, Train Acc = 0.1514, Validation Acc = 0.7987
Epoch 100: Train Loss = 0.5816, Validation Loss = 0.5833, Train Acc = 0.8419, Validation Acc = 0.8286
Epoch 200: Train Loss = 0.5168, Validation Loss = 0.5202, Train Acc = 0.8446, Validation Acc = 0.8381
Epoch 300: Train Loss = 0.4763, Validation Loss = 0.4805, Train Acc = 0.8473, Validation Acc = 0.8491
Epoch 400: Train Loss = 0.4497, Validation Loss = 0.4542, Train Acc = 0.8486, Validation Acc = 0.8553
Epoch 500: Train Loss = 0.4315, Validation Loss = 0.4362, Train Acc = 0.8506, Validation Acc = 0.8569
Epoch 600: Train Loss = 0.4187, Validation Loss = 0.4236, Train Acc = 0.8506, Validation Acc = 0.8538
Epoch 700: Train Loss = 0.4094, Validation Loss = 0.4144, Train Acc = 0.8510, Validation Acc = 0.8538
Epoch 800: Train Loss = 0.4026, Validation Loss = 0.4077, Train Acc = 0.8520, Validation Acc = 0.8522
Epoch 900: Train Loss = 0.3975, Validation Loss = 0.4026, Train Acc = 0.8527, Validation Acc = 0.8553
```

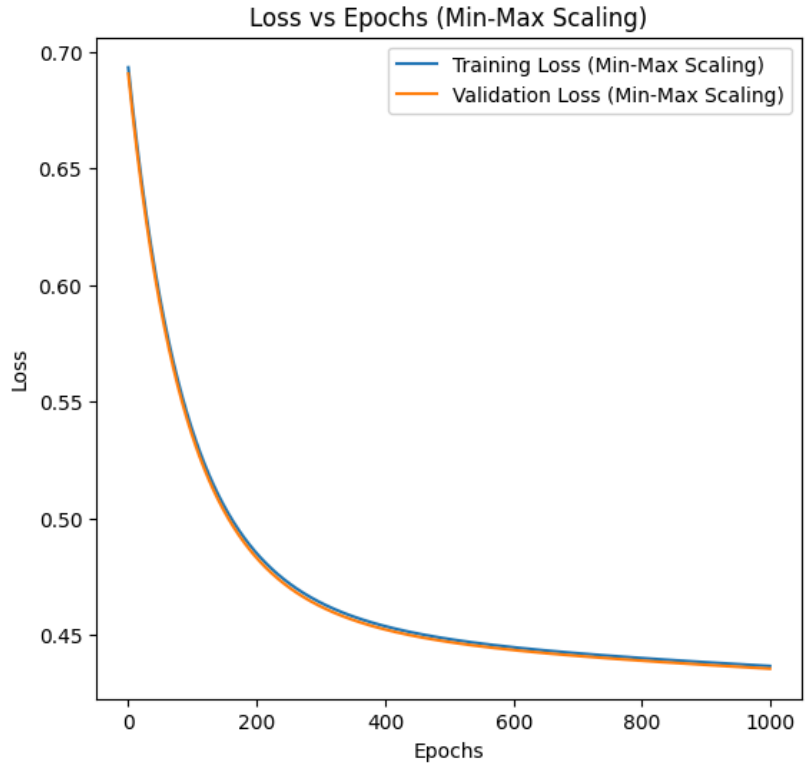




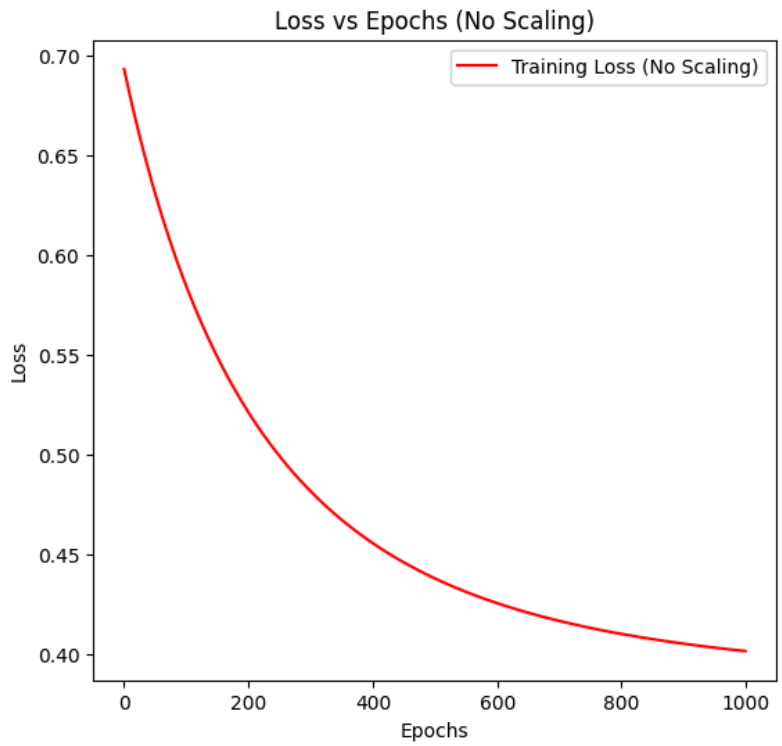
The model shows good convergence behavior as evident from both the loss and accuracy plots. The loss curves flatten out indicating that a minimum of the loss function has been reached.

The accuracy curves stabilize, which alongside the loss curves, suggests that the model parameters have adequately adapted to the underlying patterns in the data without fitting excessively to the noise or specific idiosyncrasies of the training set.

Ques-2 | Part - b



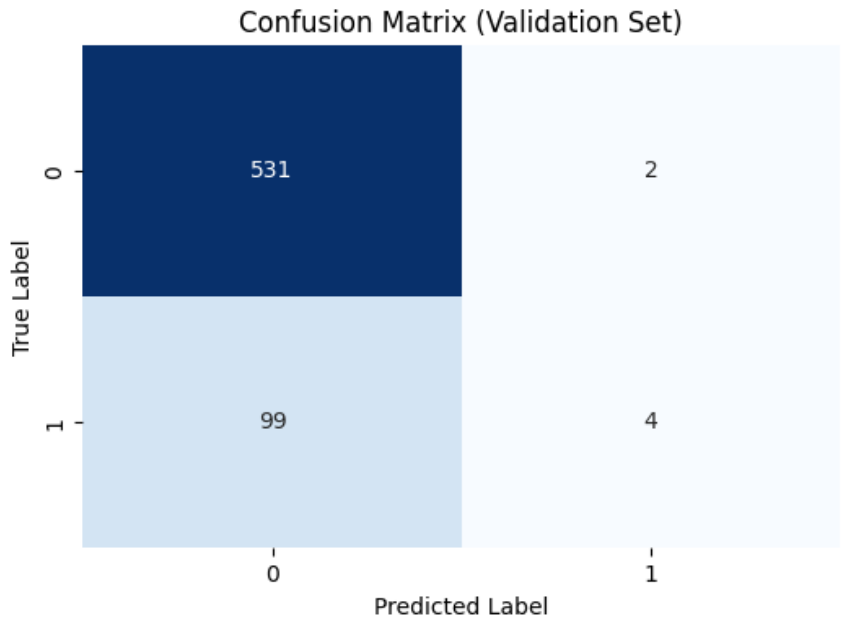
The plot demonstrates a smooth and consistent decrease in training loss over epochs, showing effective convergence even without scaling. The model successfully learns and adapts to the training data patterns, which suggests that features might be relatively uniform in scale or that the model is robust to the scale of inputs.



With Min-Max scaling, the training and validation loss both decrease smoothly and remain closely aligned throughout the training process. This indicates an efficient learning process and good generalization to validation data, supported by the uniform scaling of features.

Comparative: **No Scaling**: The model converges well, which could indicate an intrinsic balance or simplicity in the feature scales or a robustness of logistic regression to certain types of feature distribution.
Min-Max Scaling: Standardizing the range of the features to [0, 1] promotes even faster and more stable convergence, possibly by ensuring that all features contribute equally to the loss gradients and model updates

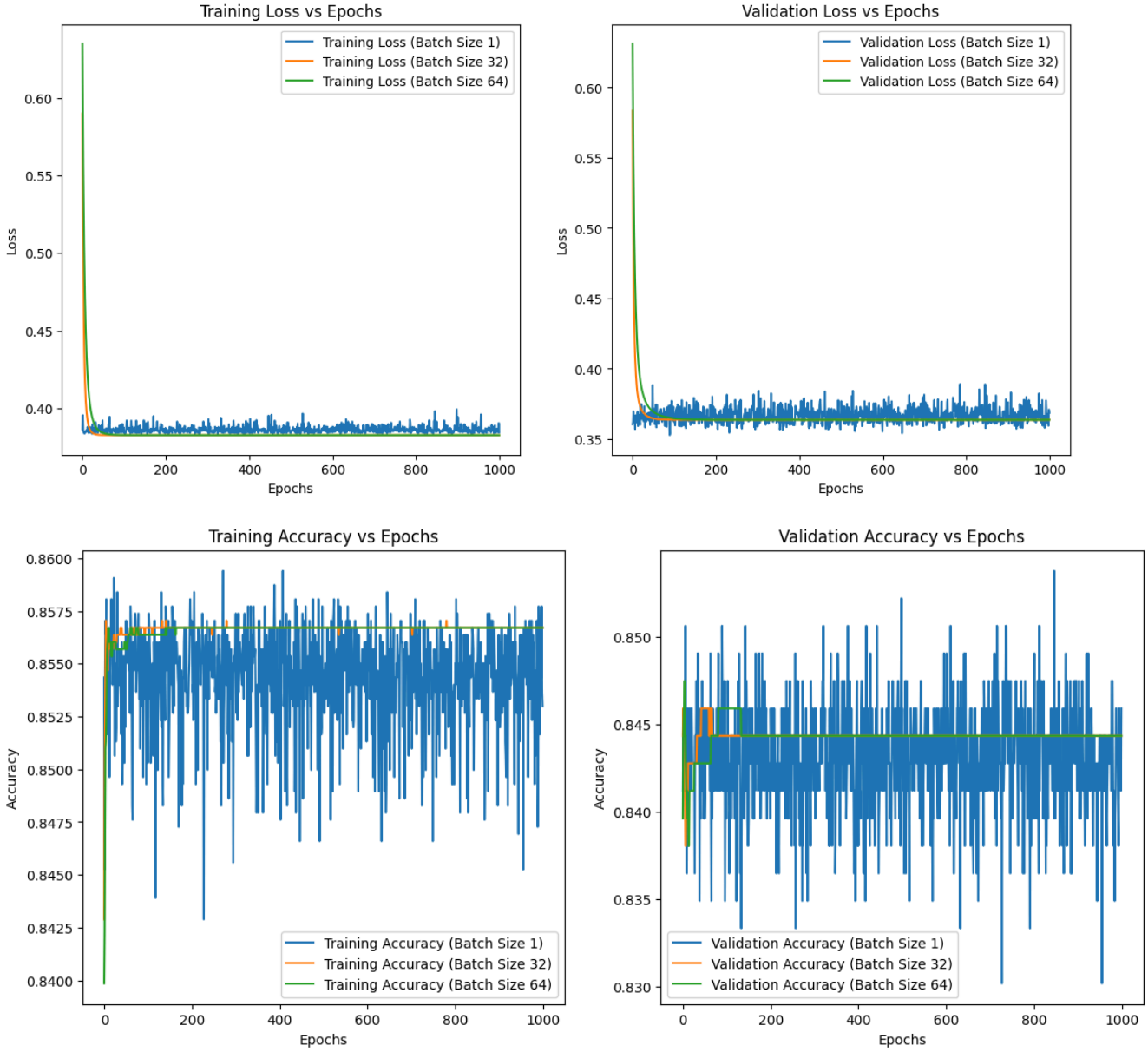
Ques-2| Part - c



Precision: 0.6667
Recall: 0.0388
F1 Score: 0.0734
ROC-AUC Score: 0.8010

The Logistic Regression model exhibits a commendable discriminative capability with a ROC-AUC score of 0.8010, indicating strong performance in differentiating between classes. The precision of 0.6667 underscores the model's reliability in making positive predictions. Enhancements in recall could further refine the model, increasing its overall effectiveness and utility in practical applications. The model has robust performance in precision and ROC-AUC where accurate positive classification is critical.

Ques-2| Part - d



The performance of Stochastic Gradient Descent (SGD) and Mini-Batch Gradient Descent with varying batch sizes shows clear distinctions in convergence stability and speed. SGD with a batch size of 1 displays more variance in training and validation accuracy, indicating rapid but unstable learning. In contrast, Mini-Batch Gradient Descent with larger batch sizes (32 and 64) exhibits smoother convergence and greater stability, albeit with a slightly slower adaptation to the training data. This analysis underscores the trade-offs between convergence speed and stability, highlighting **Mini-Batch Gradient Descent as an effective compromise** that balances efficient learning with robust performance across epochs.

Ques - 2 | Part - e

Average Accuracy: 0.8179 ± 0.0544
Average Precision: 0.5475 ± 0.3873
Average Recall: 0.0999 ± 0.1958
Average F1 Score: 0.0673 ± 0.1267

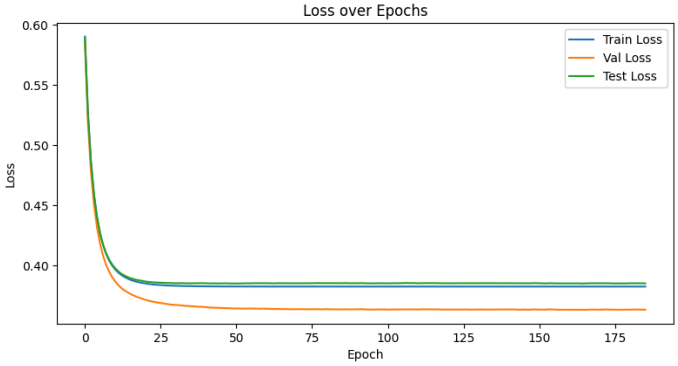
The implementation of 5-fold cross-validation reveals a solid average accuracy of 0.8179 with a modest standard deviation.
Also a decreasing trend in loss and an increasing trend in accuracy across epochs.

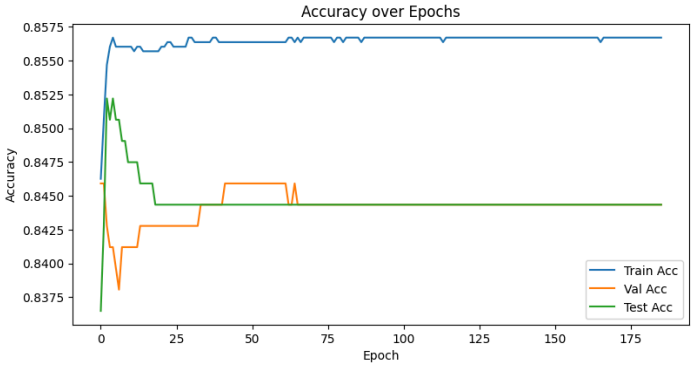
- An average accuracy of 0.8179 with a low standard deviation indicates stable and reliable model performance across different subsets.
- Moderate Precision Variability: Precision shows some variability, suggesting sensitivity to different data characteristics but maintains a reasonable average level.
- Identified Improvement Areas: Lower recall and F1 scores across folds point to areas for potential improvement in capturing positive cases, yet provide a clear focus for model enhancements.
- Robust Assessment: The methodological approach of using k-fold cross-validation ensures a thorough assessment of the model's robustness and generalization capabilities.

Ques - 2 | Part - f

Training with learning rate = 0.01, regularization = none

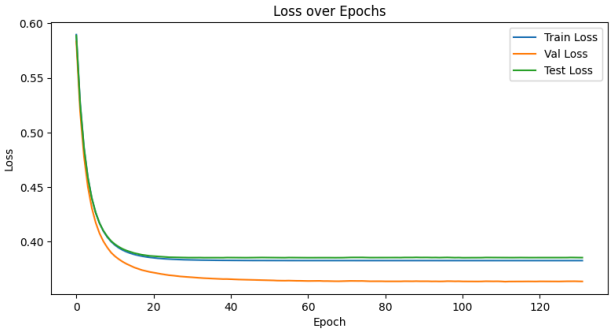
Epoch 0: Train Loss = 0.5901, Train Acc = 0.8463, Val Loss = 0.5843, Val Acc = 0.8459, Test Loss = 0.5892, Test Acc = 0.8365
Epoch 100: Train Loss = 0.3827, Train Acc = 0.8567, Val Loss = 0.3636, Val Acc = 0.8443, Test Loss = 0.3854, Test Acc = 0.8443
Early stopping triggered at epoch 185

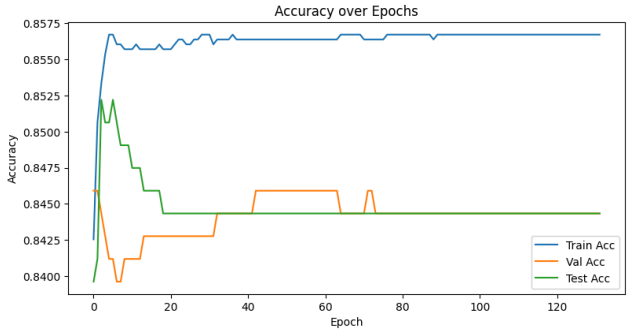




Training with learning rate = 0.01, regularization = l1

Epoch 0: Train Loss = 0.5897, Train Acc = 0.8425, Val Loss = 0.5834, Val Acc = 0.8459, Test Loss = 0.5886, Test Acc = 0.8396
Epoch 100: Train Loss = 0.3827, Train Acc = 0.8567, Val Loss = 0.3637, Val Acc = 0.8443, Test Loss = 0.3854, Test Acc = 0.8443
Early stopping triggered at epoch 131

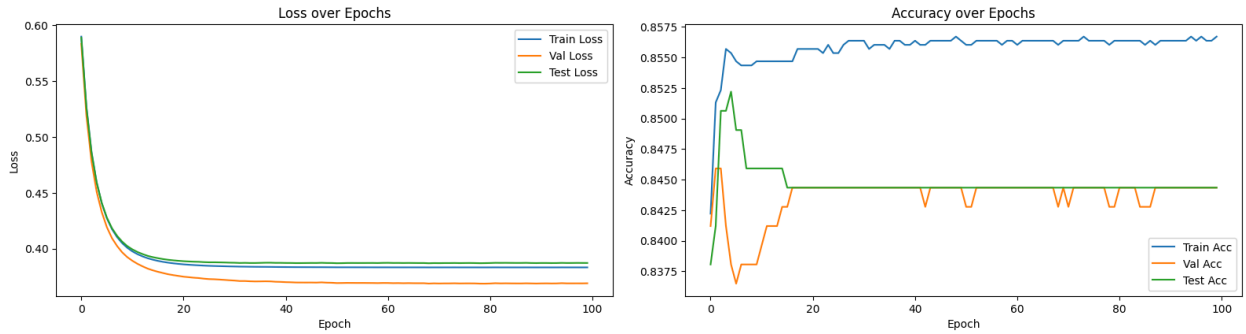




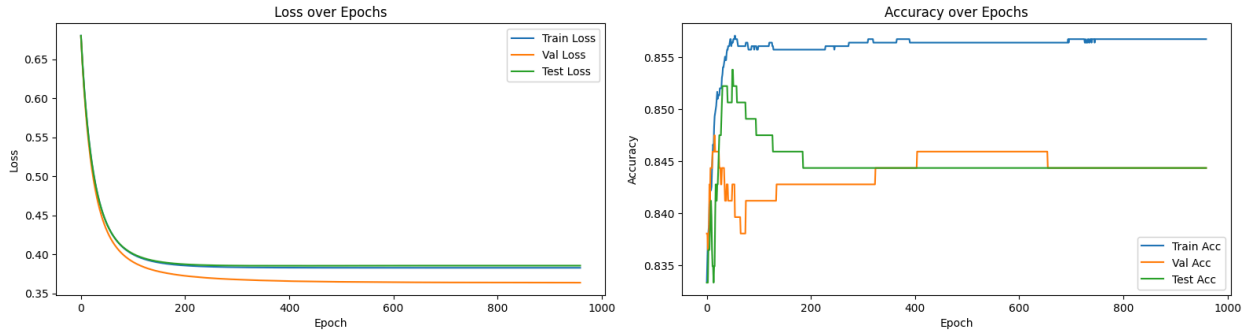
Training with learning rate = 0.01, regularization = l2

Epoch 0: Train Loss = 0.5897, Train Acc = 0.8422, Val Loss = 0.5832, Val Acc = 0.8412, Test Loss = 0.5887, Test Acc = 0.8381

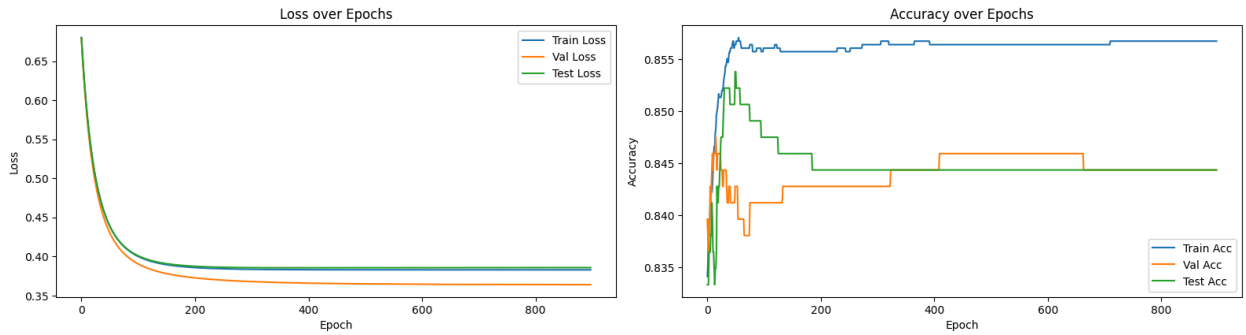
Early stopping triggered at epoch 99



Training with learning rate = 0.001, regularization = none



Training with learning rate = 0.001, regularization = l1

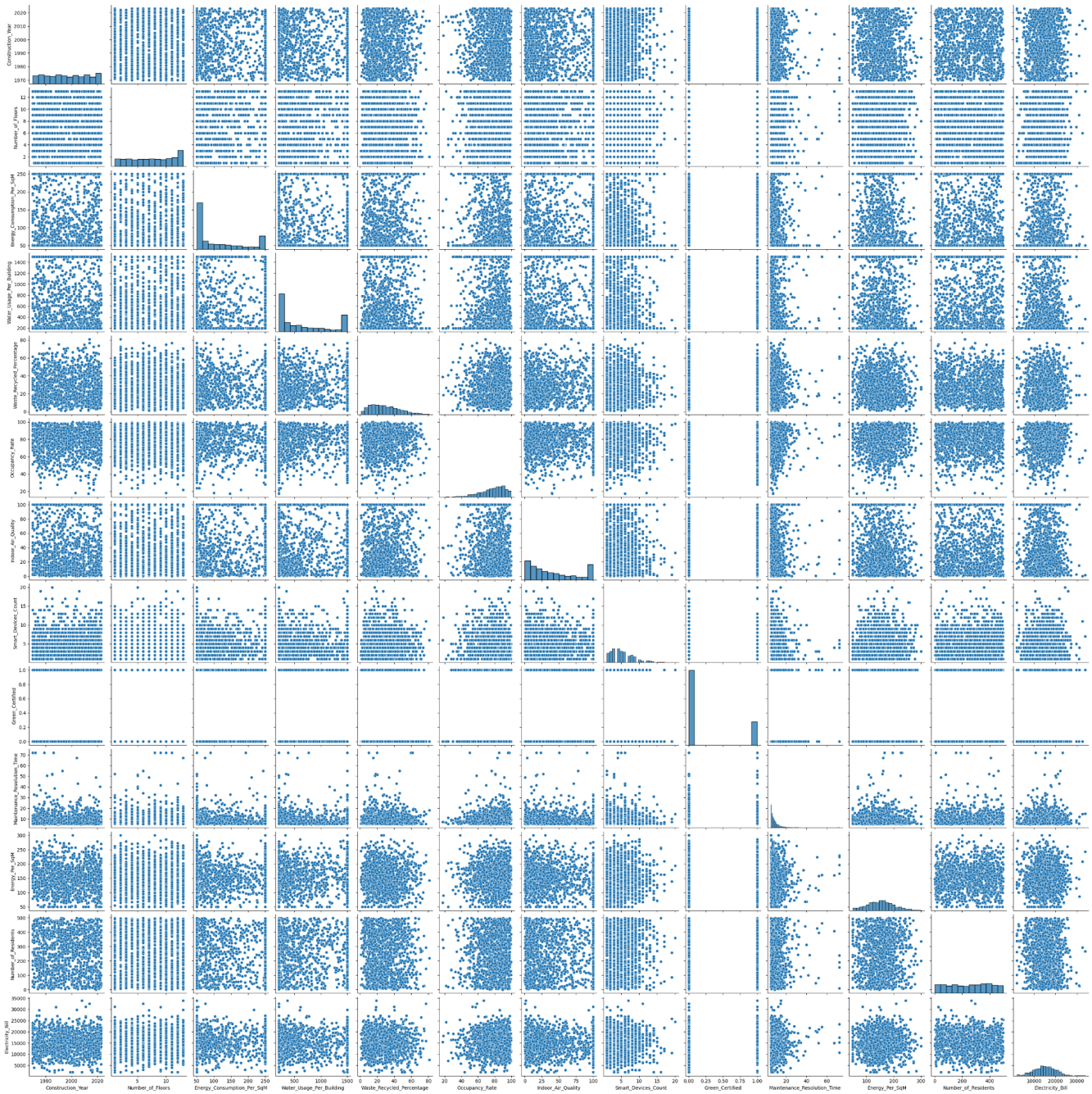


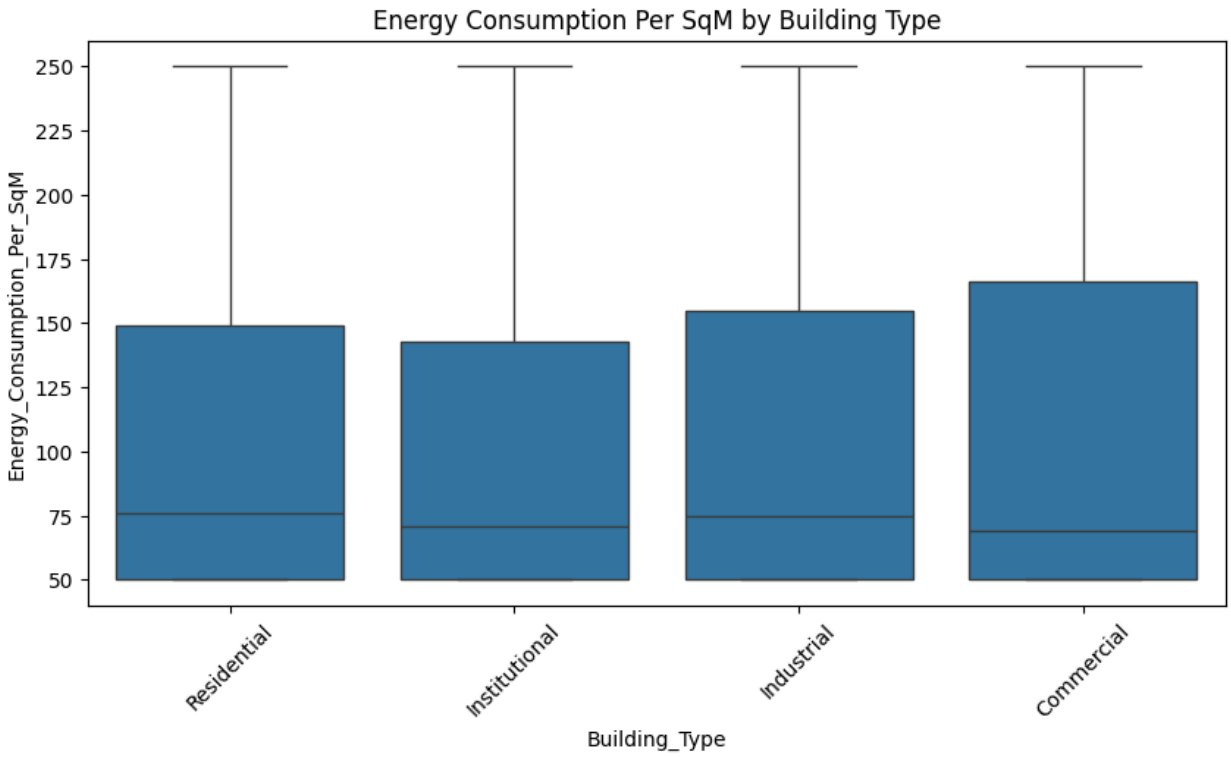
- Controlled Overfitting: Early stopping effectively controlled overfitting, maintaining consistent accuracy across validation and test sets.
- Improved Loss Trends: Loss trends show a smoother decrease, indicating better model optimization.

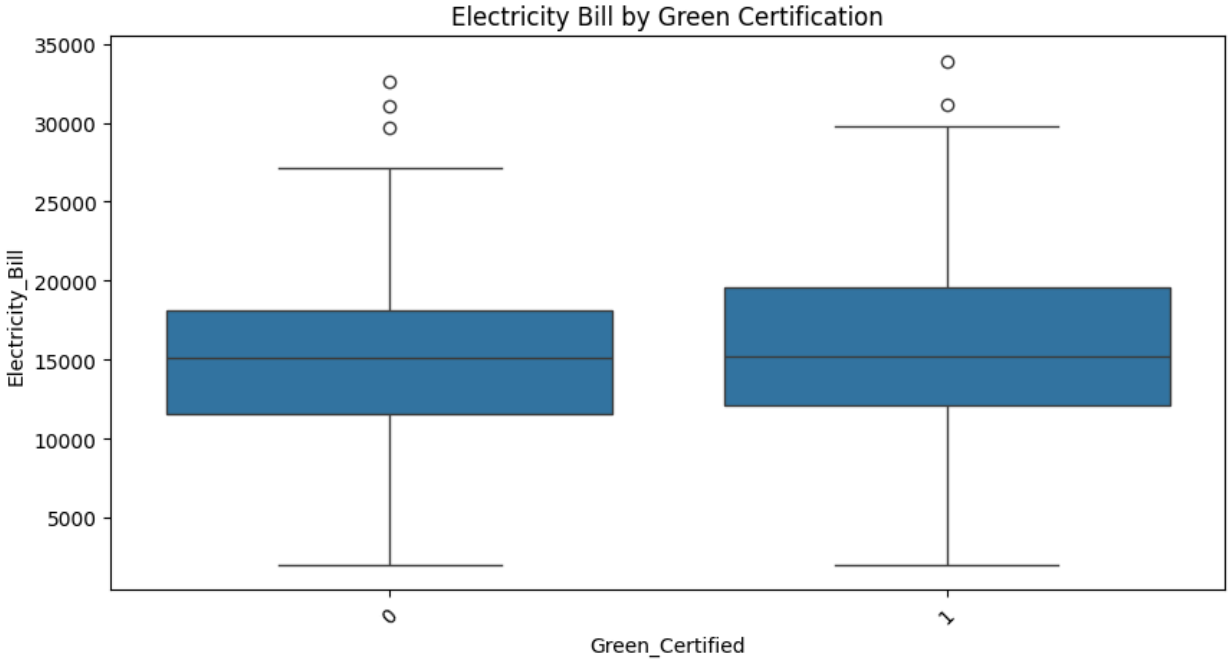
- Stable Generalization: Early stopping enhanced model generalization, evidenced by aligned accuracy metrics across different datasets.
- Effective Experimentation: Adjustments in learning rates and regularization techniques, combined with early stopping, optimized overall model performance.

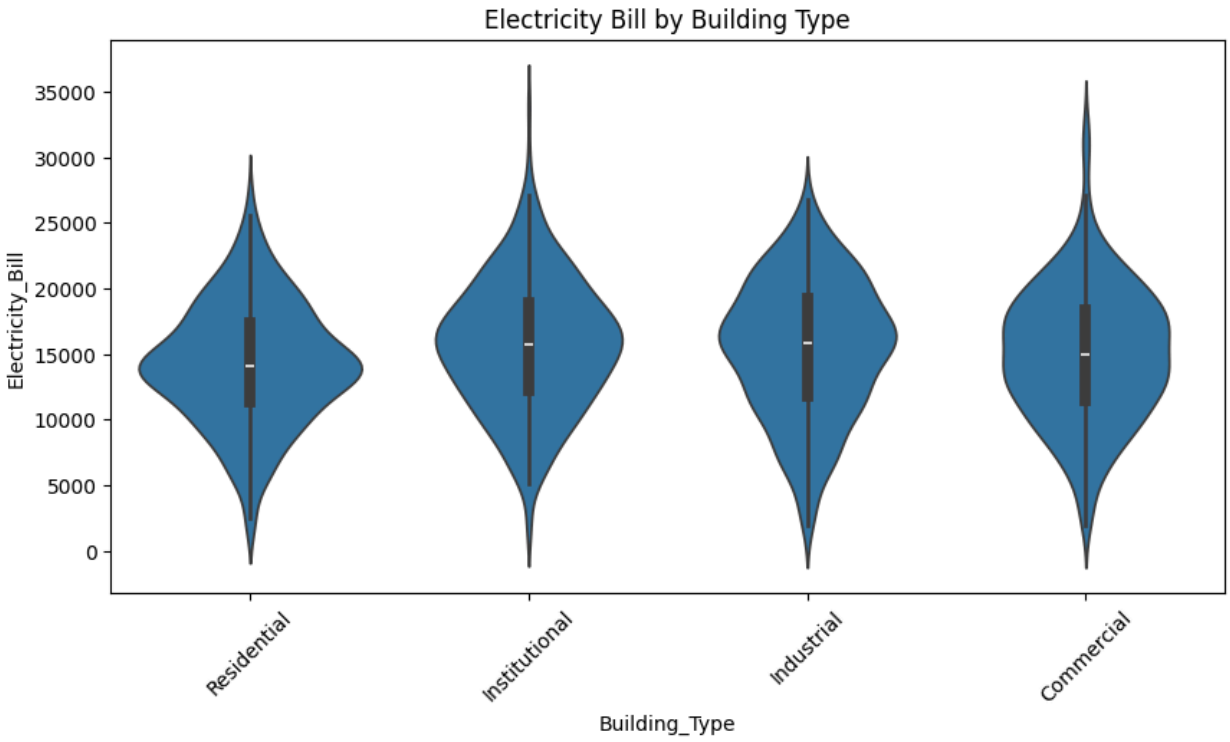
Ques - 3

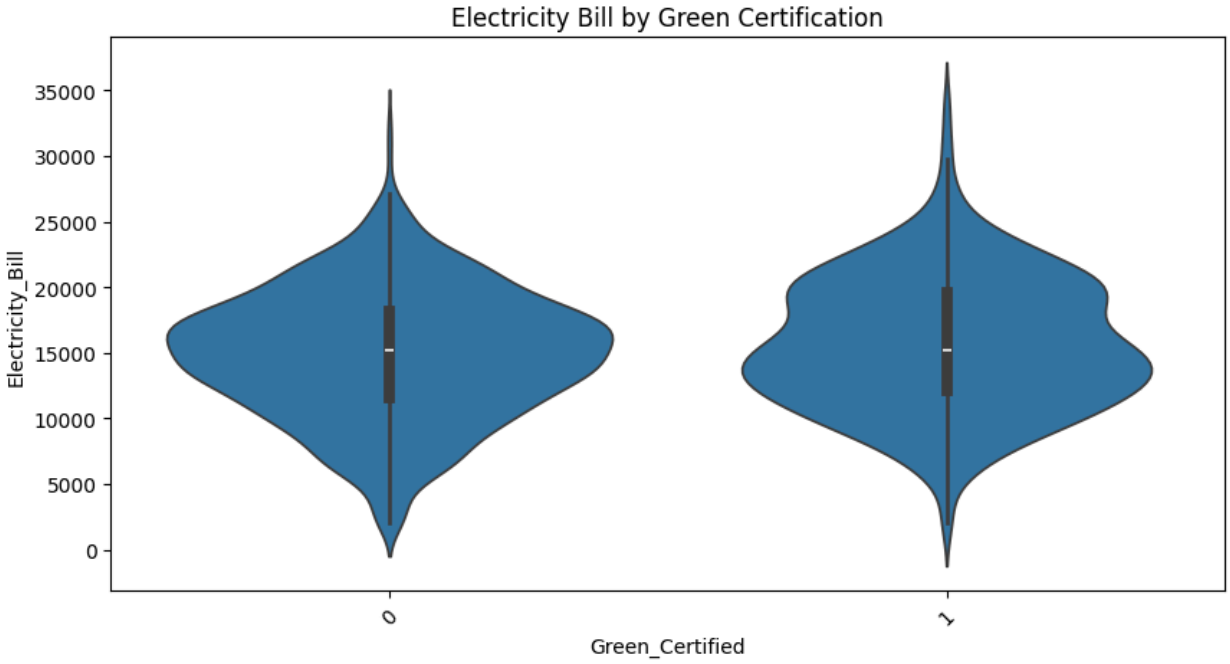
EDA:

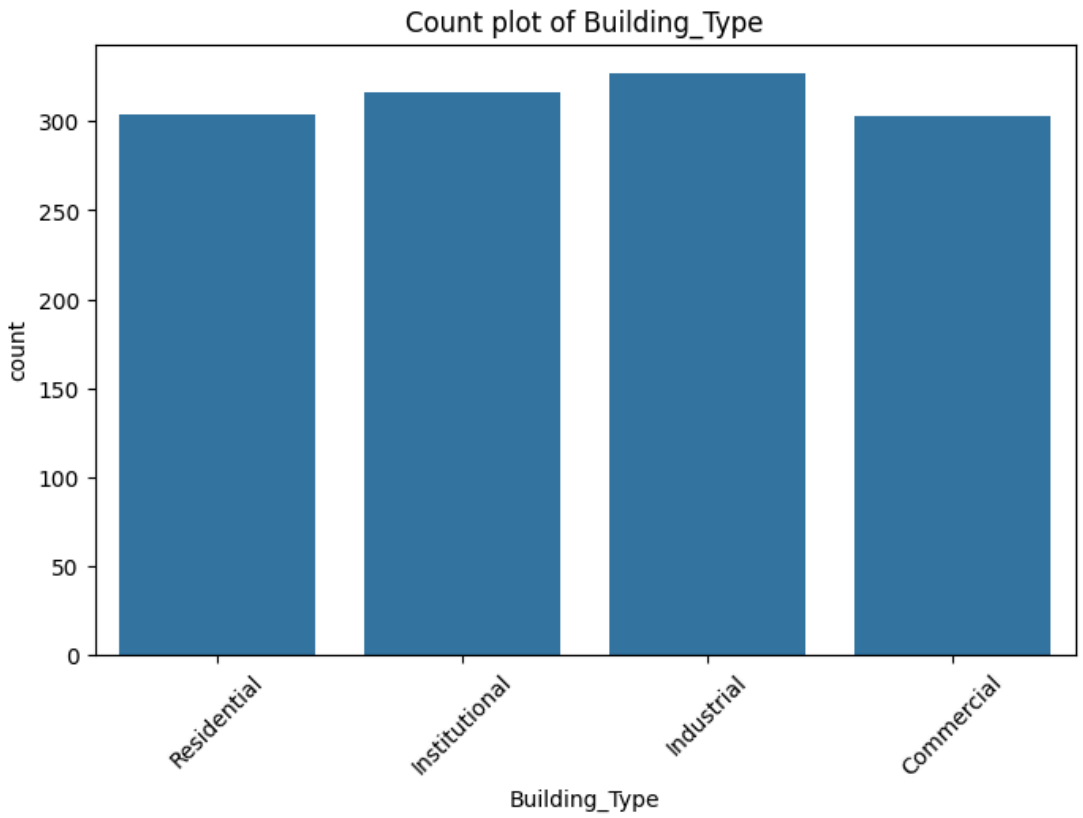

Construction_Year
Number_of_Floors
Energy_Consumption_Per_SqM
Water_Usage_Per_Building
Waste_Recycled_Percentage
Occupancy_Rate
Indoor_Air_Quality
Smart_Devices_Count
Green_Certified
Maintenance_Resolution_Time
Energy_Per_SqM
Number_of_Residents
Electricity_Bill

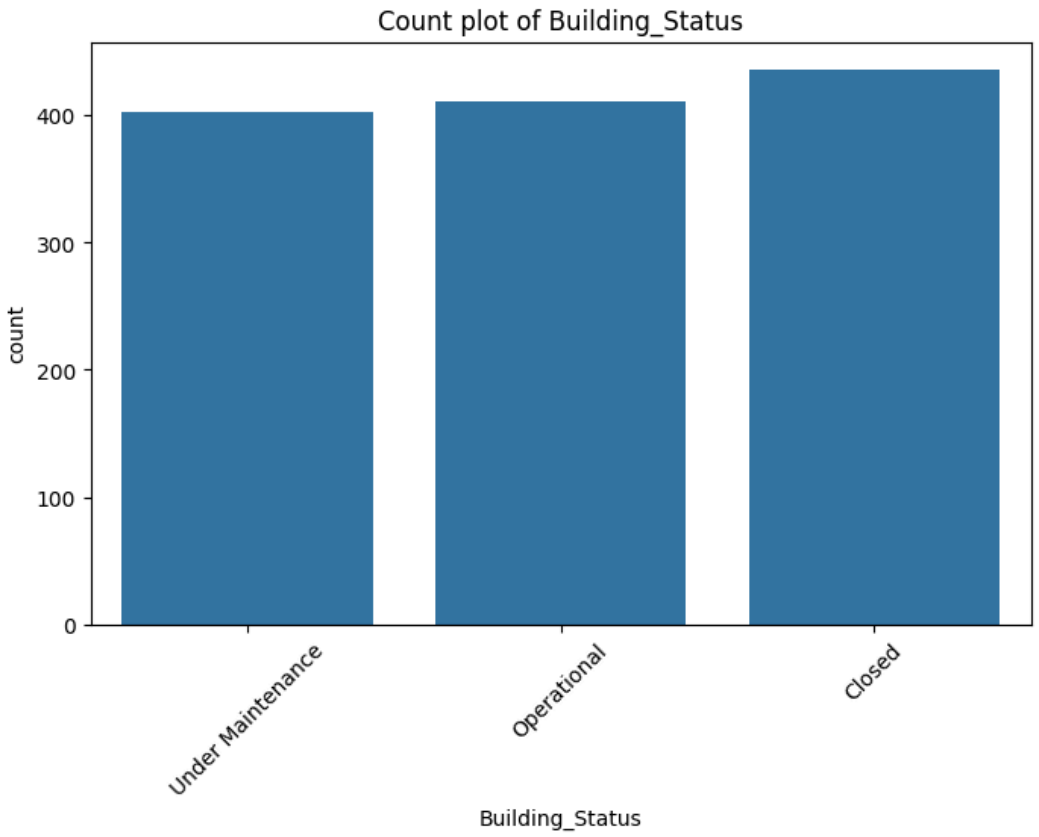

Energy Consumption Per SqM by Building Type
Energy_Consumption_Per_SqM
250
225
200
175
150
125
100
75
50
Residential
Institutional
Industrial
Commercial
Building_Type

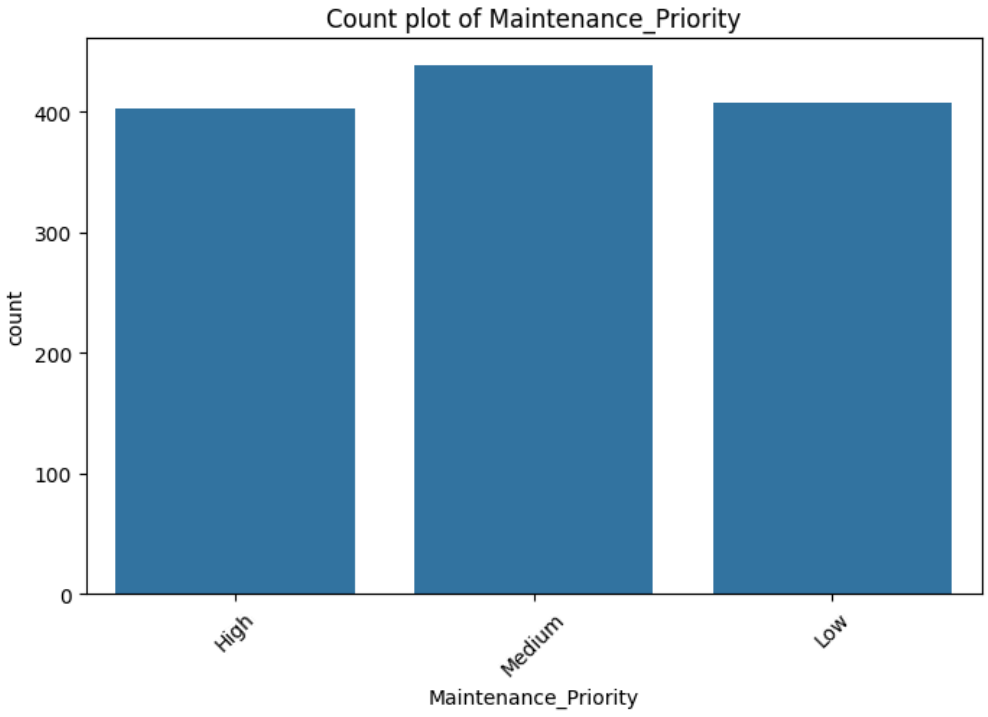

Electricity Bill by Green Certification
Electricity_Bill
35000
30000
25000
20000
15000
10000
5000
0
1
Green_Certified

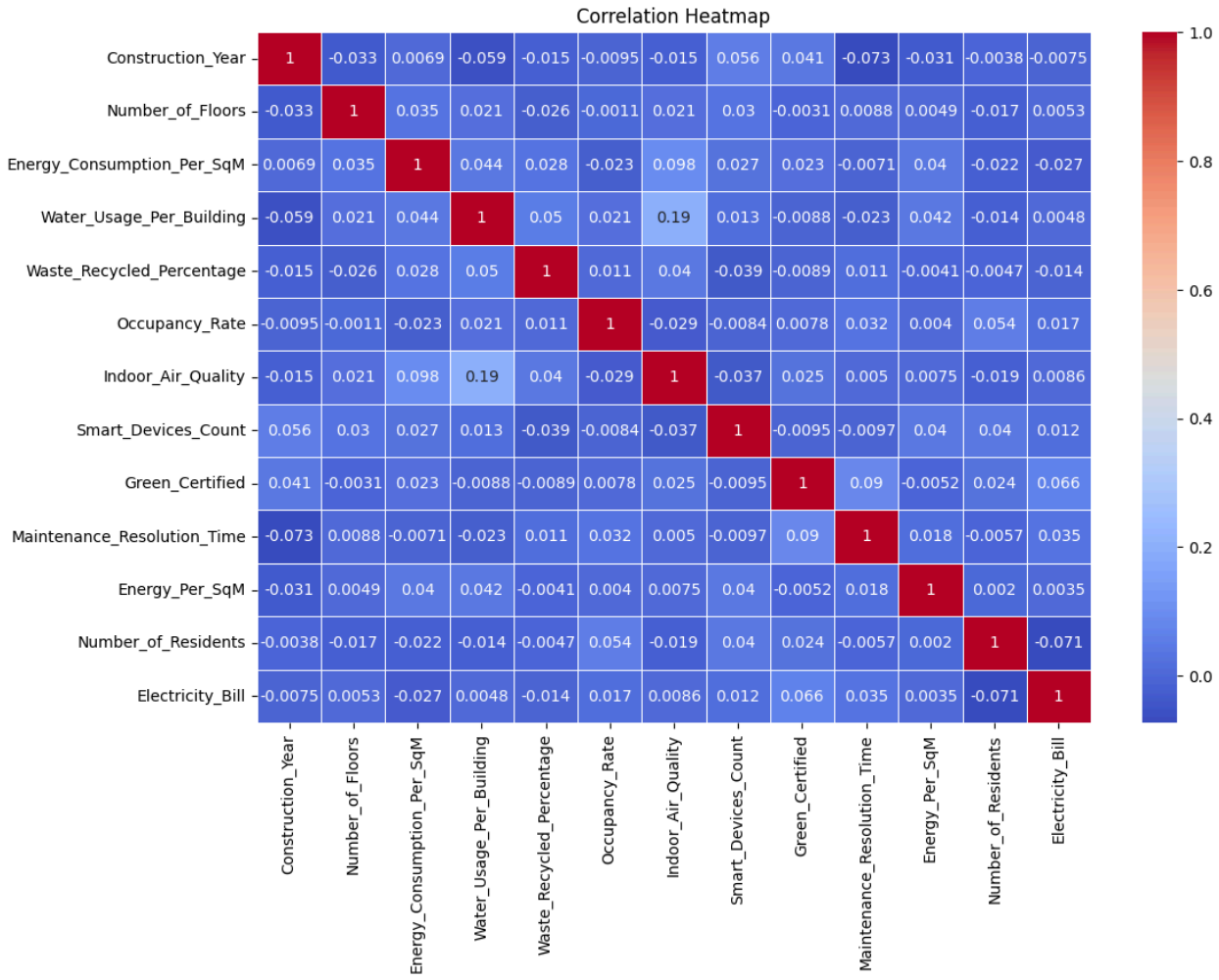

Electricity Bill by Building Type
Electricity_Bill
35000
30000
25000
20000
15000
10000
5000
0
Residential
Institutional
Industrial
Commercial
Building_Type


Electricity Bill by Green Certification
Electricity_Bill
35000
30000
25000
20000
15000
10000
5000
0
0
1
Green_Certified


Count plot of Building_Type
count
300
250
200
150
100
50
0
Residential
Institutional
Industrial
Commercial
Building_Type

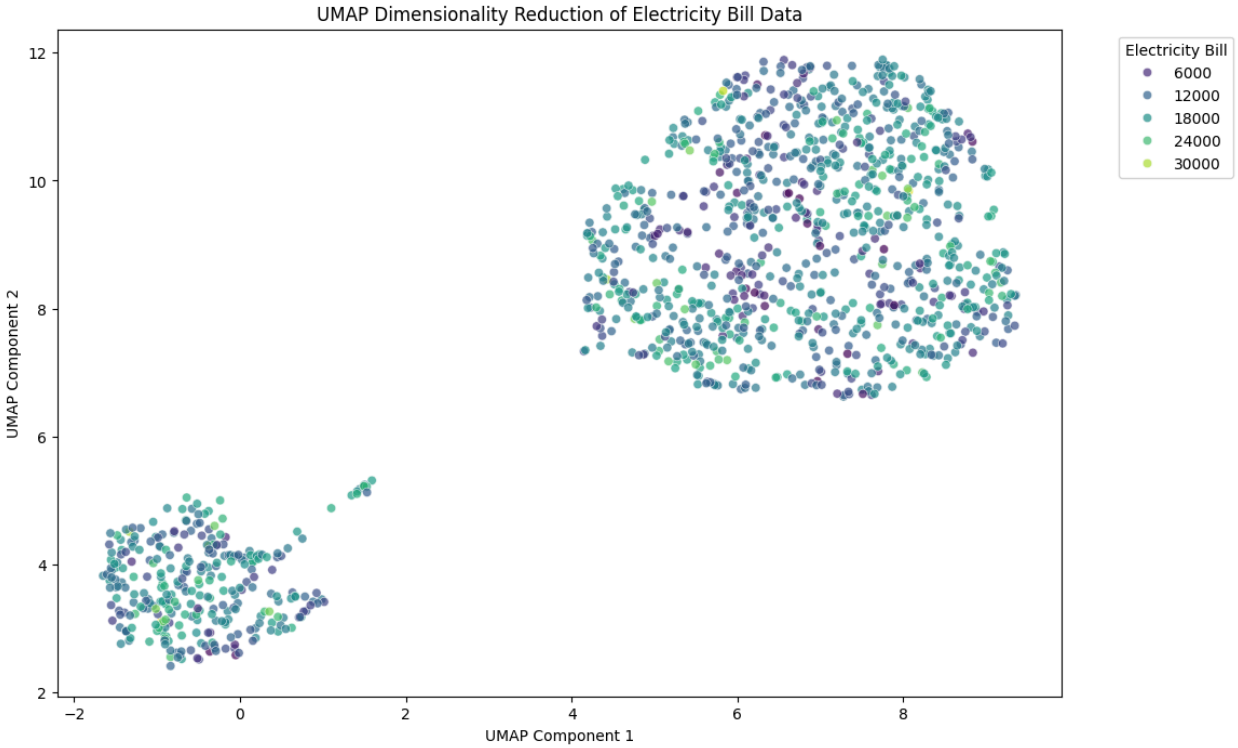

Count plot of Building_Status
count
0
100
200
300
400
Under Maintenance
Operational
Closed
Building_Status


Count plot of Maintenance_Priority
count
0
100
200
300
400
High
Medium
Low
Maintenance_Priority


Correlation Heatmap
Construction_Year 1 -0.033 0.0069 -0.059 -0.015 -0.0095 -0.015 0.056 0.041 -0.073 -0.031 -0.0038 -0.0075
Number_of_Floors -0.033 1 0.035 0.021 -0.026 -0.0011 0.021 0.03 -0.0031 0.0088 0.0049 -0.017 0.0053
Energy_Consumption_Per_SqM 0.0069 0.035 1 0.044 0.028 -0.023 0.098 0.027 0.023 -0.0071 0.04 -0.022 -0.027
Water_Usage_Per_Building -0.059 0.021 0.044 1 0.05 0.021 0.19 0.013 -0.0088 -0.023 0.042 -0.014 0.0048
Waste_Recycled_Percentage -0.015 -0.026 0.028 0.05 1 0.011 0.04 -0.039 -0.0089 0.011 -0.0041 -0.0047 -0.014
Occupancy_Rate -0.0095 -0.0011 -0.023 0.021 0.011 1 -0.029 -0.0084 0.0078 0.032 0.004 0.054 0.017
Indoor_Air_Quality -0.015 0.021 0.098 0.19 0.04 -0.029 1 -0.037 0.025 0.005 0.0075 -0.019 0.0086
Smart_Devices_Count 0.056 0.03 0.027 0.013 -0.039 -0.0084 -0.037 1 -0.0095 -0.0097 0.04 0.04 0.012
Green_Certified 0.041 -0.0031 0.023 -0.0088 -0.0089 0.0078 0.025 -0.0095 1 0.09 -0.0052 0.024 0.066
Maintenance_Resolution_Time -0.073 0.0088 -0.0071 -0.023 0.011 0.032 0.005 -0.0097 0.09 1 0.018 -0.0057 0.035
Energy_Per_SqM -0.031 0.0049 0.04 0.042 -0.0041 0.004 0.0075 0.04 -0.0052 0.018 1 0.002 0.0035
Number_of_Residents -0.0038 -0.017 -0.022 -0.014 -0.0047 0.054 -0.019 0.04 0.024 -0.0057 0.002 1 -0.071
Electricity_Bill -0.0075 0.0053 -0.027 0.0048 -0.014 0.017 0.0086 0.012 0.066 0.035 0.0035 -0.071 1
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
1.0
0.8
0.6
0.4
0.2
0.0

Part b

Umap



Part C

Train Metrics:
MSE: 24475013.1685, RMSE: 4947.2228, MAE: 4006.3285, R^2: 0.0139, Adjusted R^2: -0.0011
Test Metrics:
MSE: 24278016.1557, RMSE: 4927.2727, MAE: 3842.4093, R^2: 0.0000, Adjusted R^2: -0.0641

- Consistent Error Metrics: The consistency in MSE and RMSE between train and test sets suggests that the model is not overfitting.
- Low Mean Absolute Error: The MAE is relatively low, indicating that on average, the model's predictions are close to the actual values.
- Data Preprocessing Effectiveness: The preprocessing steps, including normalization and encoding, have prepared the data effectively for linear regression, ensuring a uniform scale for all features.

Part - D

Selected Features: ['Building_Type', 'Green_Certified', 'Building_Status']
Train Metrics with Selected Features:
MSE: 24673540.3115, RMSE: 4967.2468, MAE: 4006.7840, R^2: 0.0059, Adjusted R^2: 0.0029
Test Metrics with Selected Features:
MSE: 24181190.6472, RMSE: 4917.4374, MAE: 3825.6516, R^2: 0.0040, Adjusted R^2: -0.0081

- Error Metrics Consistency: MSE and RMSE values remained consistent, ensuring the model's stability between train and test datasets.
- MAE Reduction: Notable reduction in Mean Absolute Error (MAE) on the test dataset implies improved prediction accuracy.
- Effective Feature Selection: Utilizing RFE for feature selection highlighted key variables that marginally improved model performance.
- R^2 and Adjusted R^2 Insight: Slight variations in R^2 and Adjusted R^2 values underscore the importance of feature selection in optimizing the model's explanatory power

Part - E

Training MSE (Ridge with One-Hot Encoding): 24188934.34
Training RMSE (Ridge with One-Hot Encoding): 4918.22
Training MAE (Ridge with One-Hot Encoding): 3976.74
Training R^2 Score (Ridge with One-Hot Encoding): 0.03
Training Adjusted R^2 Score (Ridge with One-Hot Encoding): 0.01
Test MSE (Ridge with One-Hot Encoding): 24128288.50
Test RMSE (Ridge with One-Hot Encoding): 4912.06
Test MAE (Ridge with One-Hot Encoding): 3797.51
Test R^2 Score (Ridge with One-Hot Encoding): 0.01
Test Adjusted R^2 Score (Ridge with One-Hot Encoding): -0.08

- Improved Test Accuracy: Lower test MSE and MAE values indicate enhanced accuracy of the Ridge Regression model on unseen data.
- Stability Across Metrics: Consistency in MSE and RMSE between training and testing suggests stable model performance.
- Effectiveness of Encoding: One-Hot Encoding likely captured essential categorical feature nuances, slightly improving model predictiveness.
- R^2 Insight: Modest R^2 scores indicate the model's capability to explain variability in the dataset, demonstrating an improvement from baseline models.

Part - F

Results for 4 ICA components:
Train Metrics: MSE: 24691011.16422197, RMSE: 4969.005047715485, MAE: 4013.643764769286, R^2: 0.00522014517346403, Adjusted R^2: 0.0012210301791865108
Test Metrics: MSE: 24445531.025383353, RMSE: 4944.242209417269, MAE: 3852.445374277321, R^2: -0.006862255826050934, Adjusted R^2: -0.023300823268108894

Results for 5 ICA components:
Train Metrics: MSE: 24665146.333945222, RMSE: 4966.401749148494, MAE: 4013.1603260868096, R^2: 0.006262217202701148, Adjusted R^2: 0.001263536202714688
Test Metrics: MSE: 24499296.61486067, RMSE: 4949.676415167022, MAE: 3850.986633321956, R^2: -0.00907675231830285, Adjusted R^2: -0.029754554619907392

Results for 6 ICA components:
Train Metrics: MSE: 24663782.317534074, RMSE: 4966.2644228367535, MAE: 4013.2169606307066, R^2: 0.006317172264623805, Adjusted R^2: 0.0003130464172801384
Test Metrics: MSE: 24473480.235595714, RMSE: 4947.06784222692, MAE: 3847.9901113702945, R^2: -0.008013427580665367, Adjusted R^2: -0.03290264801475584

Results for 8 ICA components:
Train Metrics: MSE: 24634804.05787227, RMSE: 4963.346054616006, MAE: 4021.048300980566, R^2: 0.0074846817176699165, Adjusted R^2: -0.0005275509223487962
Test Metrics: MSE: 24610836.94980191, RMSE: 4960.931056747504, MAE: 3867.3046660507266, R^2: -0.01367087437431036, Adjusted R^2: -0.04731970007968167

- Stable Error Metrics: Consistent MSE and RMSE across different ICA component choices indicate stable error performance.
- Minor Adjustments in MAE: Marginal variations in Mean Absolute Error reflect subtle changes in model accuracy with different component numbers.
- Optimal Component Selection: The results suggest optimal component numbers might lie closer to 4 or 6, where the reduction in model error and MAE is more favorable.
- Robustness Across ICA Components: Despite the changes in the number of components, the model demonstrates robustness, maintaining a relatively consistent performance across training and testing datasets.

Part - G

- Consistent Improvement: The use of ElasticNet regularization consistently improved MSE and RMSE, especially at lower alpha values (0.1 and 0.5), indicating effective error reduction.
- Maintained Performance: The stability in MAE and R^2 across various alpha values demonstrates the model's resilience to overfitting and its capability to generalize.
- Optimal Alpha Selection: Alpha values closer to 0.1 and 0.5 yielded slightly better results, suggesting these settings as potentially optimal for balancing bias and variance.
- Robust Regularization Technique: ElasticNet, combining L1 and L2 penalties, proved robust in handling different levels of regularization, maintaining a relatively stable performance even with higher alpha values.

ElasticNet with alpha = 0.1
Test MSE: 24073398.99
Test RMSE: 4906.47
Test MAE: 3797.97
Test R^2 Score: 0.01
Test Adjusted R^2 Score: -0.07

ElasticNet with alpha = 0.5
Test MSE: 24057112.11
Test RMSE: 4904.81
Test MAE: 3803.69

Test R^2 Score: 0.01
Test Adjusted R^2 Score: -0.07

ElasticNet with alpha = 1.0
Test MSE: 24091497.90
Test RMSE: 4908.31
Test MAE: 3810.13
Test R^2 Score: 0.01
Test Adjusted R^2 Score: -0.07

ElasticNet with alpha = 5.0
Test MSE: 24233303.98
Test RMSE: 4922.73
Test MAE: 3828.87
Test R^2 Score: 0.00
Test Adjusted R^2 Score: -0.08

ElasticNet with alpha = 10.0
Test MSE: 24285856.88
Test RMSE: 4928.07
Test MAE: 3834.82
Test R^2 Score: -0.00
Test Adjusted R^2 Score: -0.08

Part - H

- Significant Reduction in Training Errors: The Gradient Boosting Regressor significantly reduced MSE and RMSE during training compared to earlier models, indicating powerful fitting capabilities.
- Improved Training Fit: A substantial R^2 Score on training data shows the model's effectiveness in capturing variance, a stark improvement over linear and regularized models.
- Challenge in Test Generalization: While the model excelled in training, it faces challenges in generalizing to test data as evidenced by negative R^2 Scores.
- Comparison Insight: The Gradient Boosting Regressor outperforms in terms of training metrics but indicates a need for adjustments or tuning to enhance test performance, possibly due to overfitting or the model’s sensitivity to the dataset's specific characteristics.

Training MSE: 15548098.78
Training RMSE: 3943.11
Training MAE: 3155.78
Training R^2 Score: 0.37
Training Adjusted R^2 Score: 0.36

Test MSE: 24900297.24
Test RMSE: 4990.02
Test MAE: 3845.57
Test R^2 Score: -0.03
Test Adjusted R^2 Score: -0.11