BUSI 651 Machine Learning Tools and Techniques

University Canada West (UCW)

CAMPUS-FALL23-66:

July 8th, 2024 - September 15th, 2024

Vancouver House: Room: E-405: On Campus

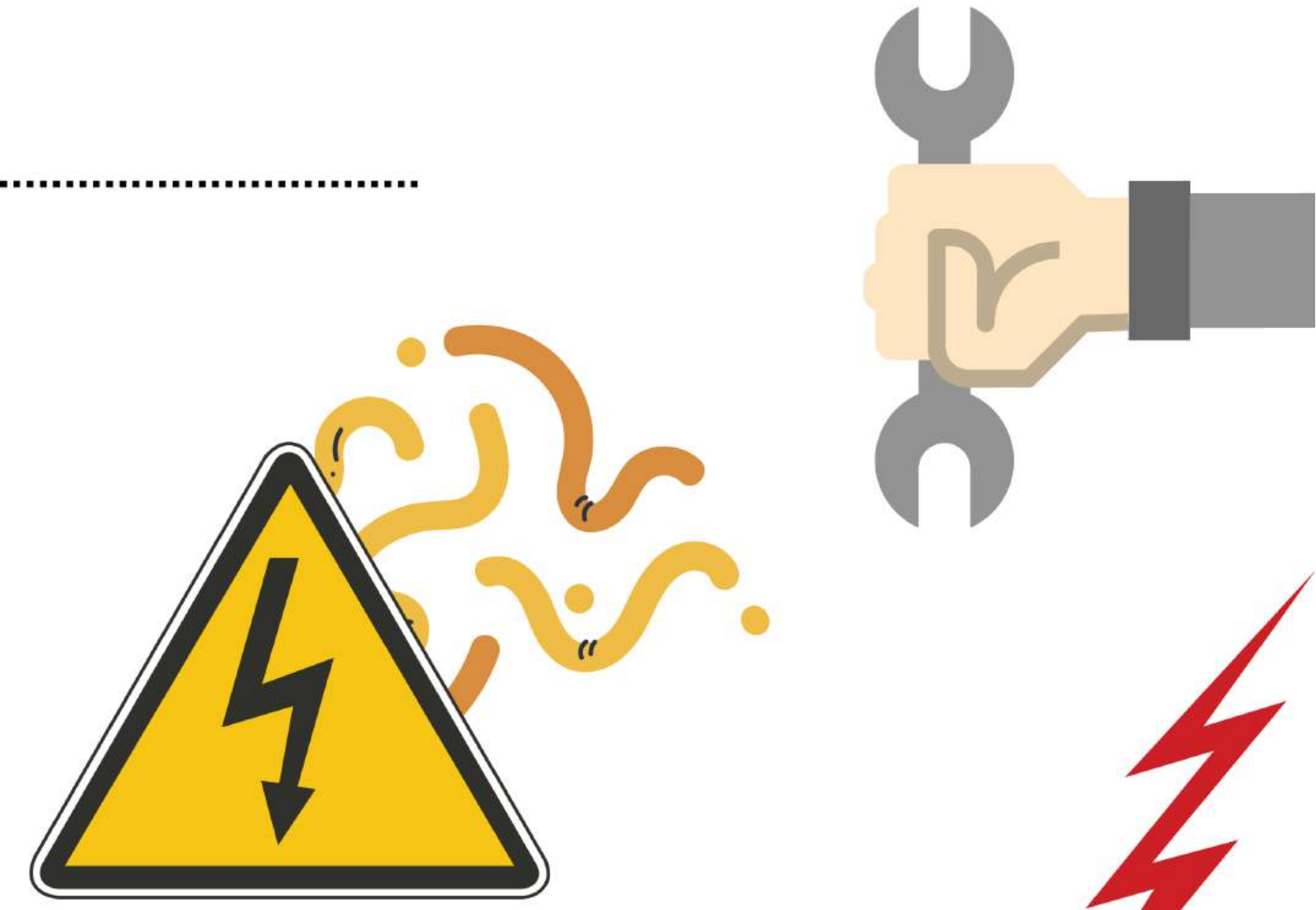
Walter Andrés Paz Callizo 2239884

Professor Sarah Gholibeigian

Due before 11:59 PM (PT) on

Sunday, August 11th, 2024..

Table of Contents



Disclaimer: This Booklet only shows answers as a brief document.
For further explanation go through the main document attached.

Executive Summary

In the following report has been carefully studied the KNN and Linear Regression concept for the further generation of new features focused mostly on the "Number of Appliances and Insulation Thickness" for the electrical construction and installing of the necessary equipment for any solutions of companies regarding the data analysis of an electrical field in the Commercial and Residential HVAC systems which is shown how these new values for a further feature can cover up for the time and the help in further worker's tasks to focus on improvement by centralizing approaches.

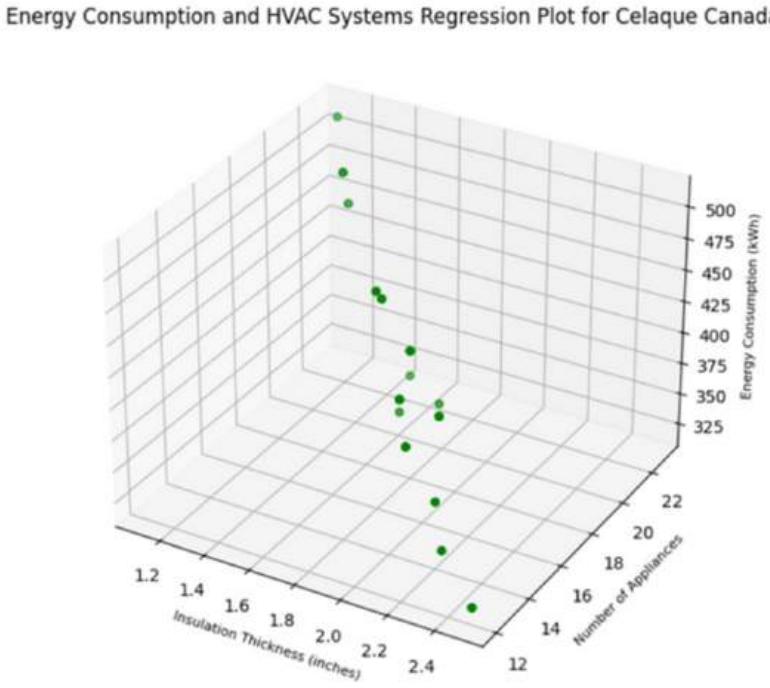
Disclaimer: This Booklet only shows answers as a brief document.
For further explanation go through the main document attached.

Introduction

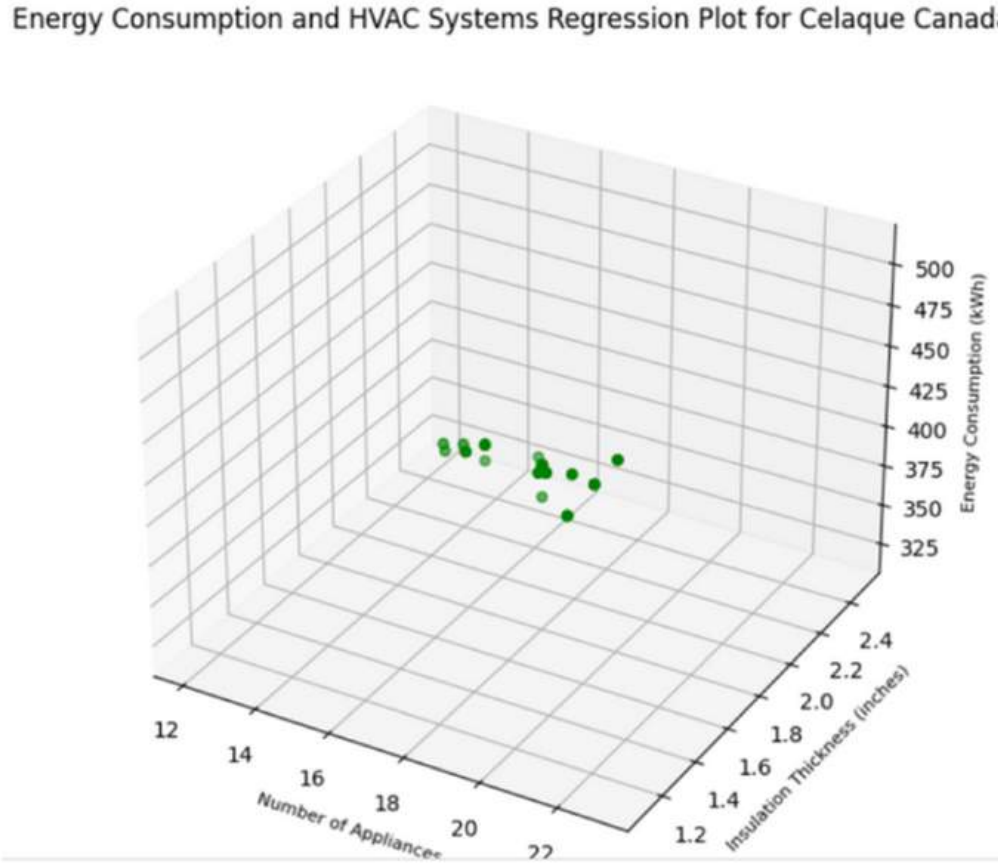
Taking care of supervised and unsupervised data with the process of orientating a Machine Learning Algorithm for the superlative construction of a business decision through the usage of Machine Learning and the interpretation for the future outcomes such as new features mixing up the most influential variables among the correct correlation based on the proximity towards the line of Regression and the Variance encountered inside of the model while experiencing with less energy consumption in construction. This is the main reason a contractor is needing to check on various features for the construction and installation of electrical equipment to a customer's building or residential homes in the data bases of the construction company.

Disclaimer: This Booklet only shows answers as a brief document.
For further explanation go through the main document attached.

Results and Investigation

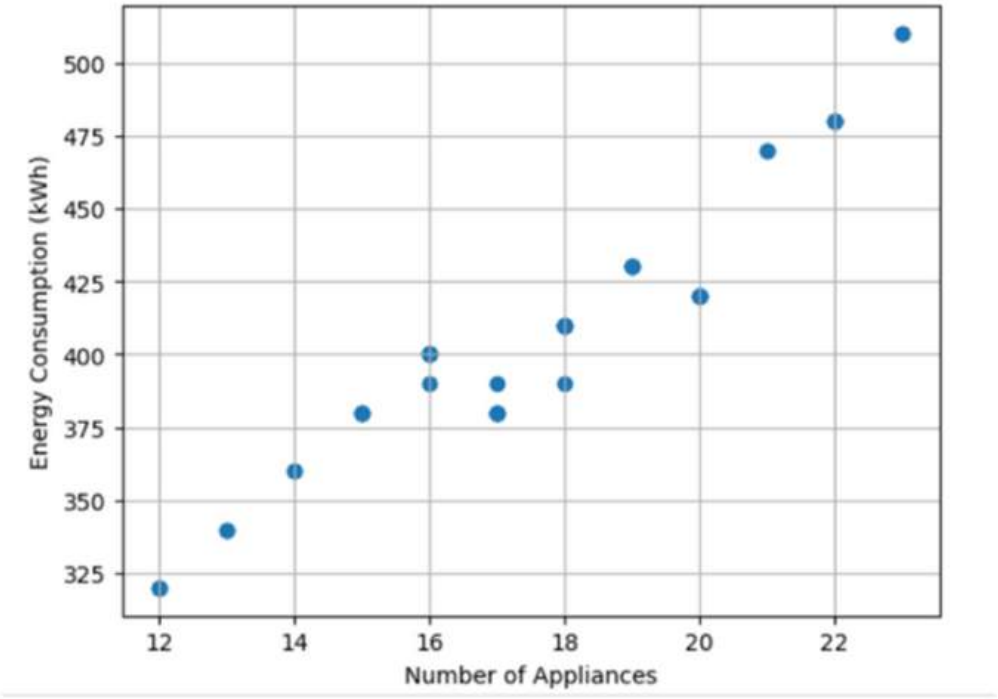


Energy Consumption and HVAC Systems Regression Plot for Insulation Thickness (Inches) and Number of Appliances (NOA)

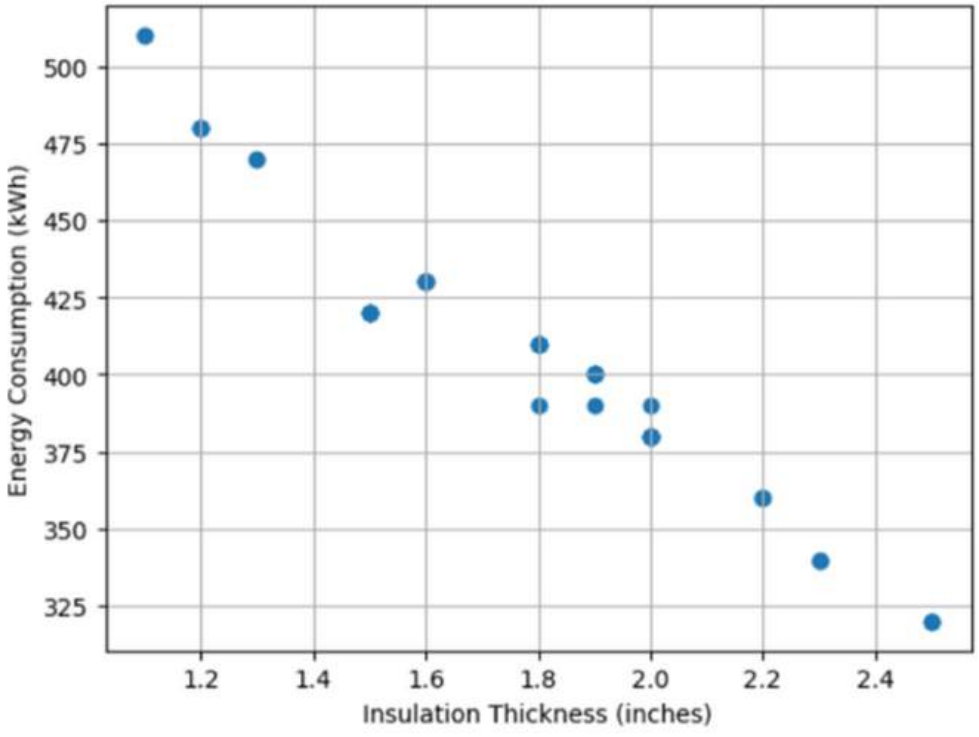


Disclaimer: This Booklet only shows answers as a brief document.
For further explanation go through the main document attached.

Results and Investigation



Energy Consumption and Number of Appliances compared towards Linear Regression (LR)



Energy Consumption and Insulation Thickness (Inches) compared towards Linear Regression (LR)

Disclaimer: This Booklet only shows answers as a brief document.
For further explanation go through the main document attached.

Results and Investigation

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).
Index(['Room Area (sq. ft.)', 'Number of Appliances',
       'Outside Temperature (C)', 'Insulation Thickness (inches)',
       'Building Type', 'HVAC System',
       'Average Temperature in last 24 hours (C)', 'Energy Consumption (kWh)'],
      dtype='object')
Missing values in X_train:
Room Area (sq. ft.)                        0
Number of Appliances                       0
Outside Temperature (C)                    0
Insulation Thickness (inches)              0
Average Temperature in last 24 hours (C)   0
Building Type_ Residential                 0
HVAC System_ Split AC                      0
HVAC System_ Window AC                     0
dtype: int64
Missing values in y_train:
0
Data types in X_train:
Room Area (sq. ft.)                        int64
Number of Appliances                       int64
Outside Temperature (C)                    int64
Insulation Thickness (inches)            float64
Average Temperature in last 24 hours (C)   int64
Building Type_ Residential                  bool
HVAC System_ Split AC                       bool
HVAC System_ Window AC                      bool
dtype: object
Shape of X_train: (45, 8)
Shape of y_train: (45,)
▾ LinearRegression
LinearRegression()
```

Linear Regression (LR) preparation coming from the Preparing Data and Testing Data

```
Mean Absolute Error: 4.774496874760634
Mean Squared Error: 35.69230892691998
R² Score: 0.9833601331667251
Predicted Energy Consumption: [289.59893118]
R² Score (Accuracy): 0.9833601331667251
```

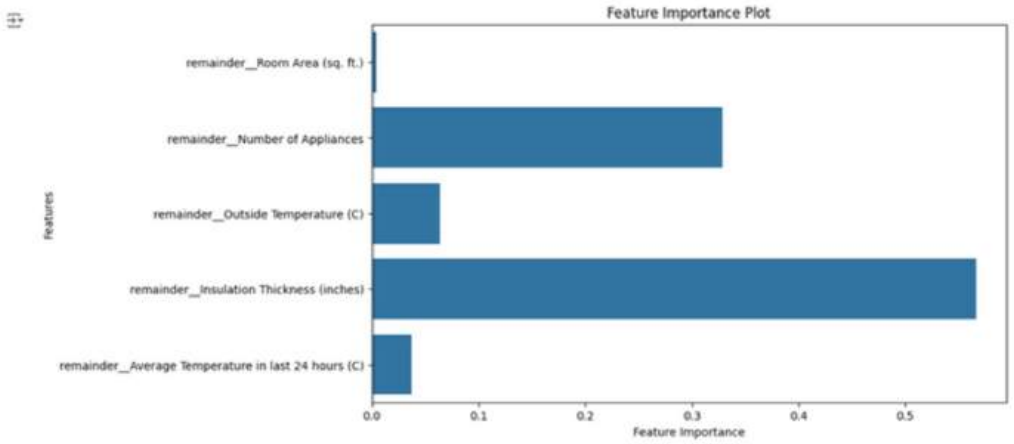
Accuracy of the Default Model (DM)

```
Drive already mounted at /content/drive; to attempt
Column names: Index(['Room Area (sq. ft.)', 'Number
       'Outside Temperature (C)', 'Insulation Thick
       'Building Type', 'HVAC System',
       'Average Temperature in last 24 hours (C)',
      dtype='object')
Missing values in X_train:
Room Area (sq. ft.)                        0
Number of Appliances                       0
Outside Temperature (C)                    0
Insulation Thickness (inches)              0
Building Type                              0
HVAC System                                0
Average Temperature in last 24 hours (C)   0
dtype: int64
Missing values in y_train:
0
Mean Absolute Error: 5.120067441936206
Mean Squared Error: 40.35978293380985
R² Score: 0.9811841420846844
Predicted Energy Consumption: [408.1270997]
/usr/local/lib/python3.10/dist-packages/sklearn/pre
  warnings.warn(
```

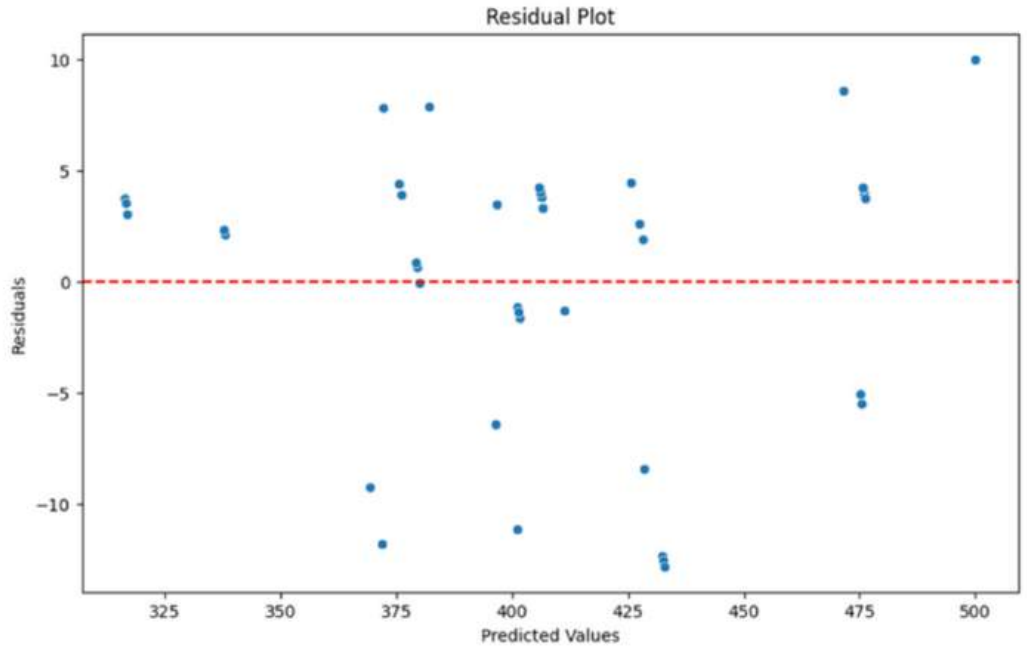
Accuracy of the Point 1

Disclaimer: This Booklet only shows answers as a brief document.
For further explanation go through the main document attached.

Results and Investigation



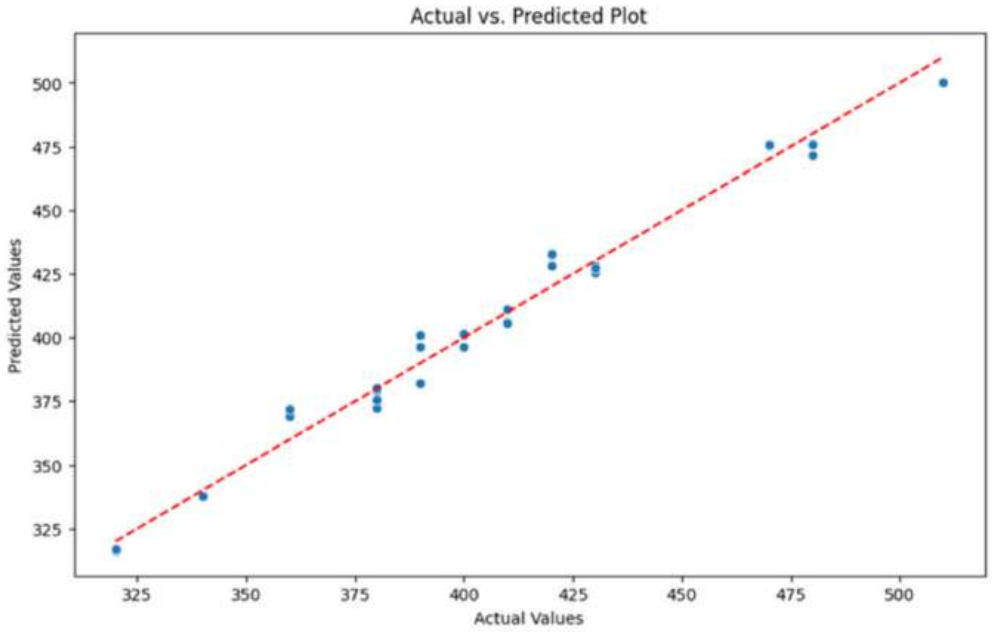
Feature Importance Point 1 Comparison



Residual Plot Point 1 Predicted Values vs Residuals

Disclaimer: This Booklet only shows answers as a brief document.
For further explanation go through the main document attached.

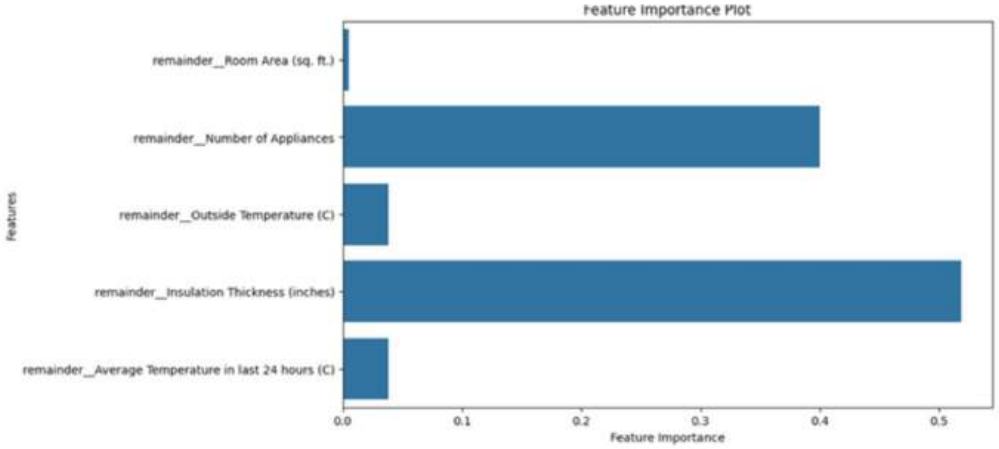
Results and Investigation



Actual vs Predicted Plot Point 1

```
Missing values in y_train:
0
Mean Absolute Error: 5.120067441936206
Mean Squared Error: 40.35978293380985
R² Score: 0.9811841420846844
Predicted Energy Consumption: [443.68701886]
/usr/local/lib/python3.10/dist-packages/sklearn
  warnings.warn(
```

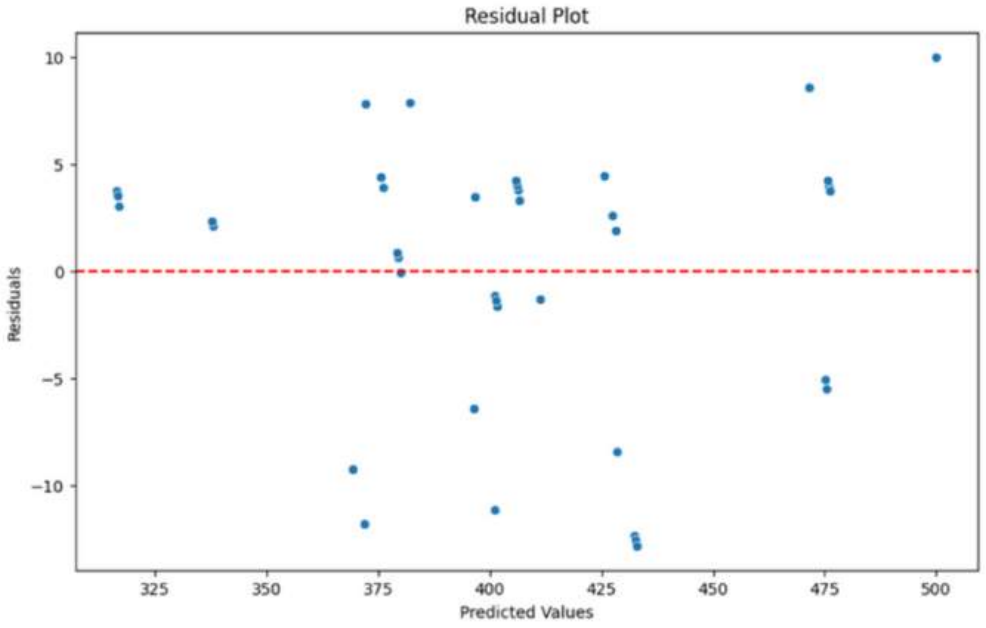
98.11% Accuracy in Point 2



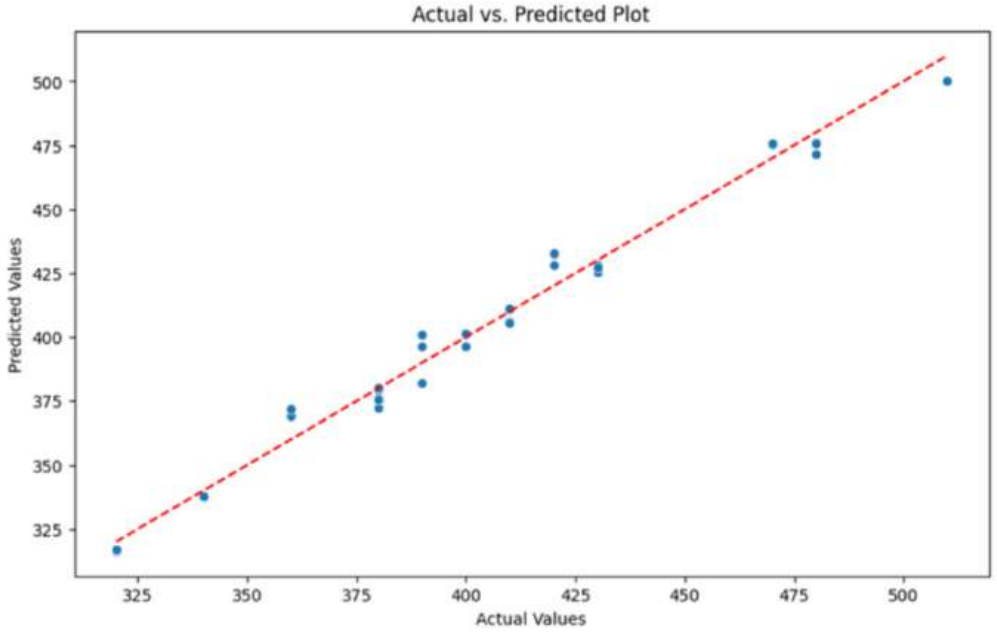
Feature Importance Plot Point 2

Disclaimer: This Booklet only shows answers as a brief document.
For further explanation go through the main document attached.

Results and Investigation



Residual Plot of Predicted vs Residuals in Point 2



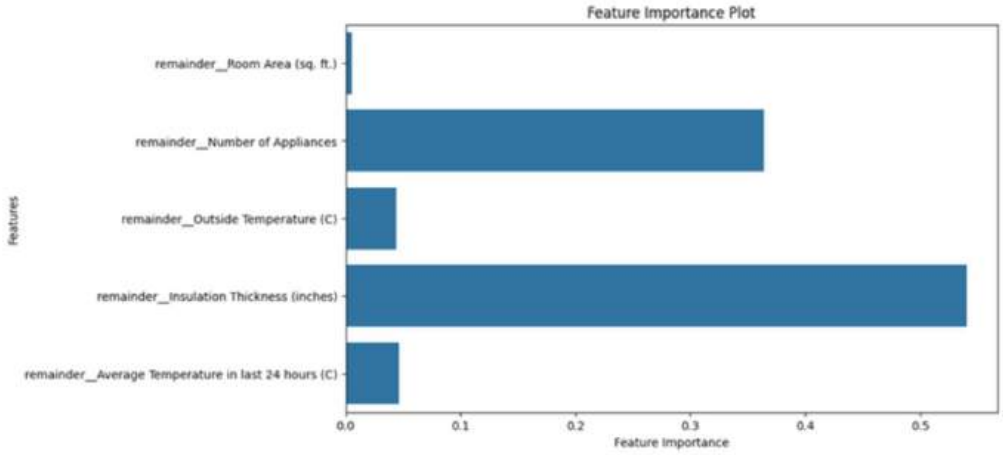
Predicted vs Actual Values in Point 2

Disclaimer: This Booklet only shows answers as a brief document.
For further explanation go through the main document attached.

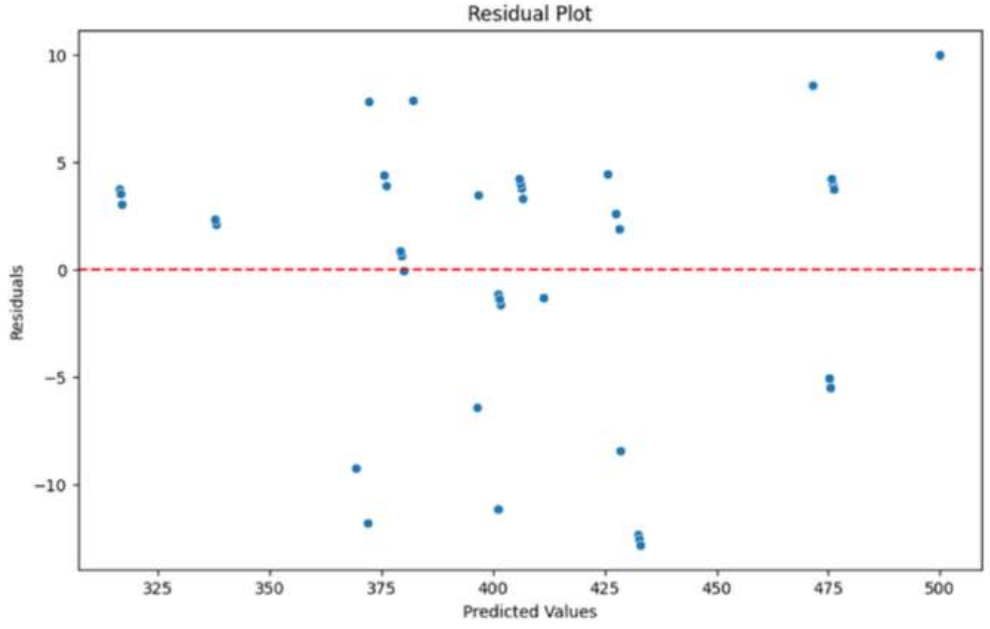
Results and Investigation

```
Mean Absolute Error: 5.120067441936206
Mean Squared Error: 40.35978293380985
R² Score: 0.9811841420846844
Predicted Energy Consumption: [361.01811883]
/usr/local/lib/python3.10/dist-packages/sklearn/pr
  warnings.warn(
```

98.11% Accuracy in Point 3



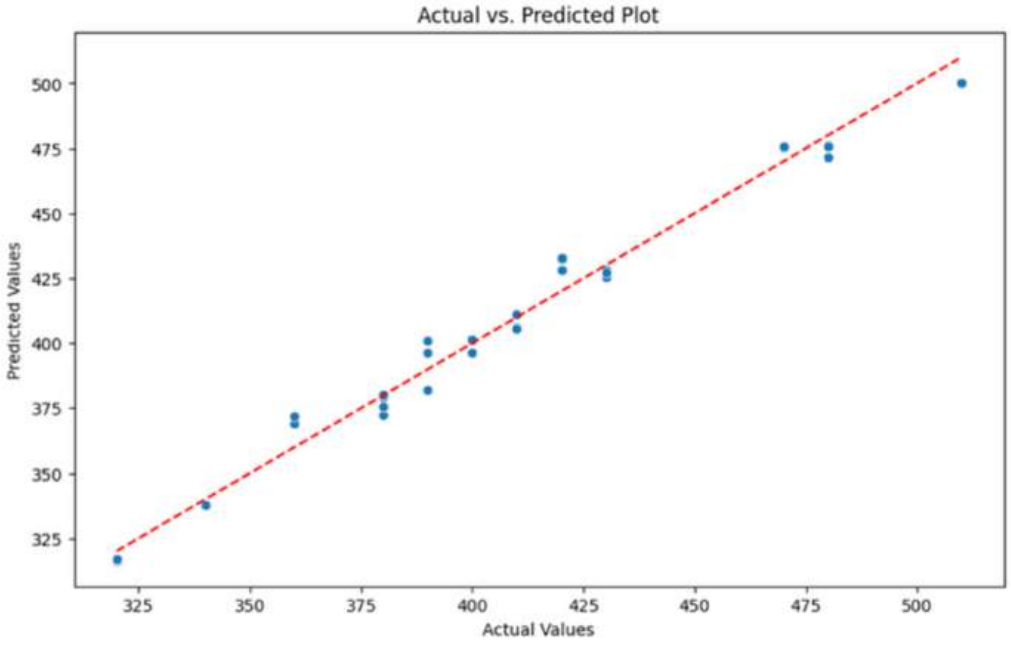
Feature Importance Plot Point 3



Residuals vs Predicted Values Point 3

Disclaimer: This Booklet only shows answers as a brief document.
For further explanation go through the main document attached.

Results and Investigation



Predicted vs Actual Values Plot 3

```
###MEAN SQUARED ERROR###
```

```
from sklearn.metrics import mean_squared_error
# Assuming y_train holds your actual target valu
print(mean_squared_error(y_train, y_pred))
```

```
40.35978293380985
```

```
from sklearn.metrics import mean_squared_error
y_true = [385, 425, 350]  # Ground truth values
y_pred = [408.1270997, 443.68701886, 361.0181188

# Compute MSE
mse = mean_squared_error(y_true, y_pred)

print(f'Mean Squared Error: {mse}')
```

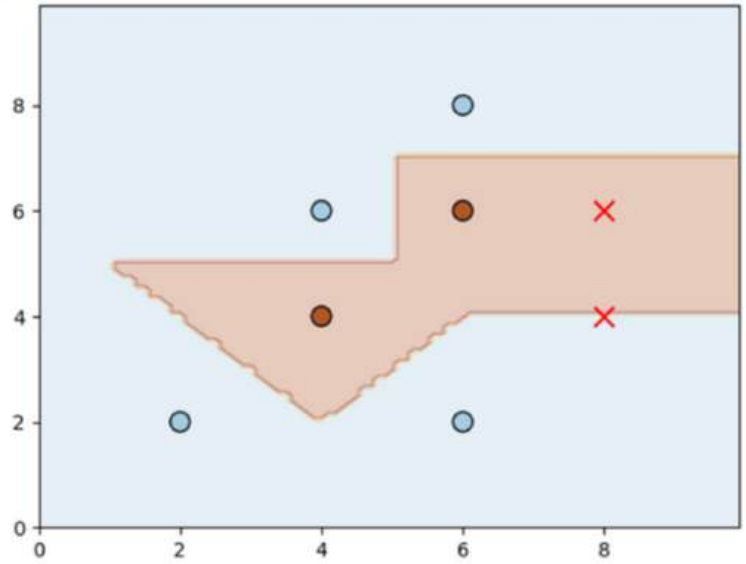
```
Mean Squared Error: 335.15545231991194
```

Mean Squared Error (MAE) of selected values

Disclaimer: This Booklet only shows answers as a brief document.
For further explanation go through the main document attached.

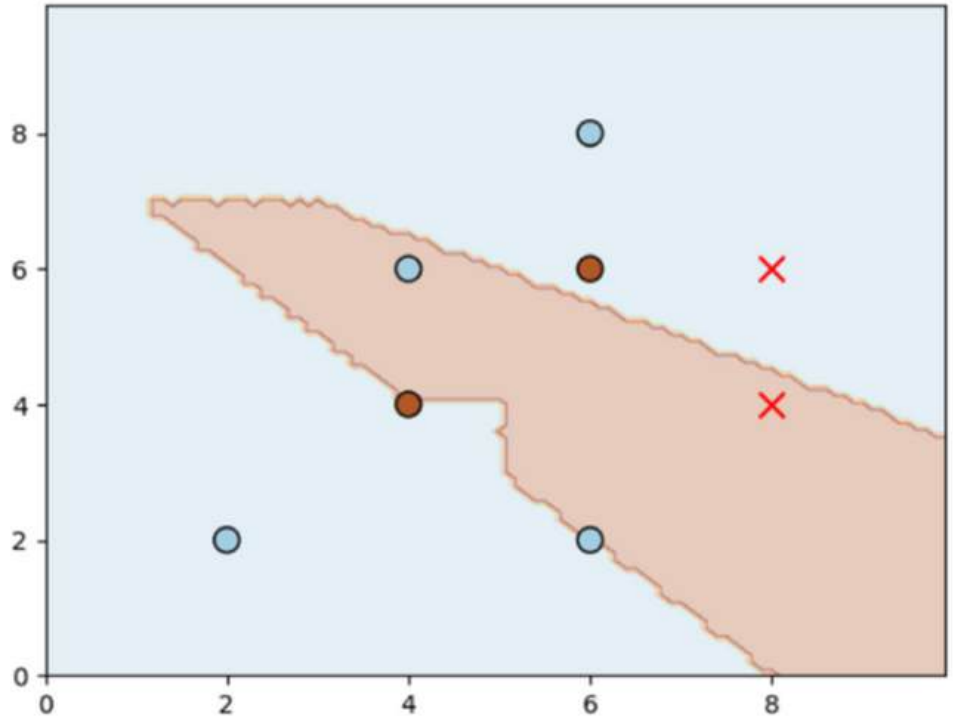
Results and Investigation

```
<ipython-input-41-7f8d5a4ddd1e>:23: UserWarning: You passed a edgecolor/edgecolors ('k') for a
  plt.scatter([8], [6], c='red', edgecolors='k', marker='x', s=100)  # point (8, 6)
<ipython-input-41-7f8d5a4ddd1e>:24: UserWarning: You passed a edgecolor/edgecolors ('k') for a
  plt.scatter([8], [4], c='red', edgecolors='k', marker='x', s=100)  # point (8, 4)
```



KNN for Question II with K1

```
<ipython-input-43-7a8c206ca8b8>:23: UserWarning: You passed a edgecol
  plt.scatter([8], [6], c='red', edgecolors='k', marker='x', s=100)
<ipython-input-43-7a8c206ca8b8>:24: UserWarning: You passed a edgecol
  plt.scatter([8], [4], c='red', edgecolors='k', marker='x', s=100)
```



KNN with K3 for Question II

Disclaimer: This Booklet only shows answers as a brief document.
For further explanation go through the main document attached.

Results and Investigation

```
Accuracy: 0.00%
              precision    recall  f1-score   support

           +       0.00      0.00      0.00       1.0
           -       0.00      0.00      0.00       1.0

    accuracy                           0.00       2.0
   macro avg       0.00      0.00      0.00       2.0
weighted avg       0.00      0.00      0.00       2.0

Predictions for new points: ['+' '-']
Point [7.81 5.33] is classified as +
Point [9.43 5.29] is classified as -
```

Points used for the Classification

```
# Create the dataset
data = {
    'X1': [8.27, 1.58, 5.92, 9.44, 2.11, 4.71, 3.82, 6.98, 3.15, 8.9, 7.65, 9.83, 1.94, 7.13, 5.77, 4.36, 5.09, 3.42, 2.76, 9.6],
    'X2': [5.59, 5.87, 5.87, 5.83, 5.57, 5.94, 5.84, 5.91, 5.42, 5.94, 5.77, 5.29, 5.36, 5.28, 5.47, 5.31, 5.65, 5.24, 5.71, 5.52],
    'Y': ['+', '-', '-', '+', '-', '+', '+', '-', '-', '-', '+', '-', '-', '-', '-', '+', '-', '+', '+', '-']
}

df = pd.DataFrame(data)
X = df[['X1', 'X2']].values
y = df['Y'].values

# Normalize the features
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# Define the points to predict
new_points = np.array([[7.81, 5.33], [9.43, 5.29]])
new_points_scaled = scaler.transform(new_points)

# Initialize NearestNeighbors model
neighbors = NearestNeighbors(n_neighbors=3)
neighbors.fit(X_scaled)

# Find the indices of the closest neighbors for each new point
distances, indices = neighbors.kneighbors(new_points_scaled)

print(f'Indices of the closest neighbors for each point: {indices}')
```

```
Indices of the closest neighbors for each point: [[13 11 14]
 [11 13 19]]
```

Index Generation for KNN

Disclaimer: This Booklet only shows answers as a brief document.
For further explanation go through the main document attached.

Results and Investigation

Question 1: After the procedure of multiple Linear Regression's (LR's) inside of the Collaboratory Notebook for Jupiter.

a) Primary feature: Insulation Thickness **Secondary feature:** Number of Appliances.

Reason: In both the feature of Number of Appliances and Insulation Thickness not only share better relationship with the Energy Consumption after handling the data with Multiple -Linear Regression but it also conducts a better prediction for a more precise analysis through the Regression Model prepared on the data.

b) Feature not contributing: Type of Building and the HVAC System (HVACS).

Mitigation strategy: As an analyst is better to either take it out of the possible variable analysis as these variables or features are merely categorical and will not influence the Linear Regression (LR) as values can only be interpreted in binary code or Boolean Values (1 and 0's) which is better only to try with Numerical Values and not the Categorical Values. Categorical Values are only used for True and False Tables.

If no such feature exists, justify your claim: -----.

Disclaimer: This Booklet only shows answers as a brief document. For further explanation go through the main document attached.

Results and Investigation

c) Apply multiple linear regression:

Energy consumption for point 1: 408.127. **Energy consumption for point 2:** 443.678. **Energy consumption for point 3:** 361.018.

The energy consumption for the Points 1, 2 and 3 are the following: For Point 1 the energy consumption of 408.127, in Point 2 the energy consumption predicted is of 443.678 and an energy consumption of 361.018.

d) Mean Squared Error (MSE) regression loss: The Mean Squared (MSE) Regression Loss is of **40.359** within Point 1, Point 2 and Point 3 out of the process of complex crossing Multiple-Linear Regression (MLR) with new points for analysis.

Disclaimer: This Booklet only shows answers as a brief document.
For further explanation go through the main document attached.

Results and Investigation

e) Recommendation to include the new feature (with reason): A new feature should include certain considerations from the **Insulation Thickness (IT) and the Number of Electrical Appliances** due to a stable alignment in Linear Regression (LR) delivering a closer linear fitting inside of the diagrams for the three-dimensional comparison of X, Y and Z features. Additionally, the residuals are closer to the Regression Line (RL) signaling a further co-relation of variables through variance and the variation in data explained by the R-Squared furthermore inside of the code with values close to one or one hundred percentage of relation.

However, an essential feature to consider inside arguments due to a better control in any electrical and construction projects can be the need of an Investment Risk Assessment as financials have a huge impact in decision-making from the business and customer's needs, identification for risks, performance, planning and time management for the building built.

Disclaimer: This Booklet only shows answers as a brief document.
For further explanation go through the main document attached.

Results and Investigation

Question 2: After KNN modeling with one and three Nearest Neighbors

Decision Boundary for K=1: In the Decision Boundary for K=1 the new area covered by the K-Nearest Neighbors which is the following:

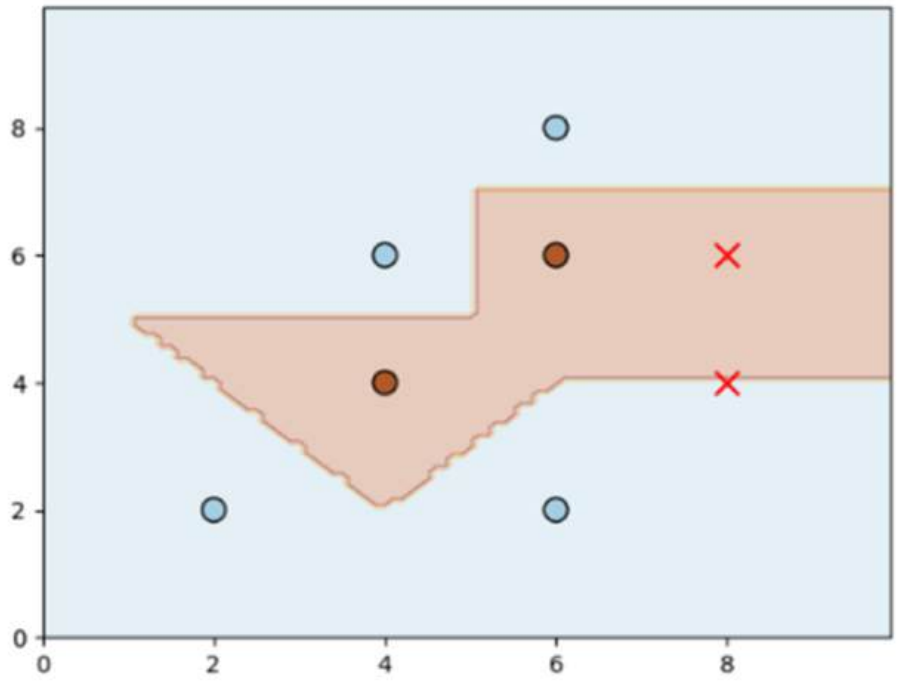
```
Accuracy: 0.96
Precision: 0.96
Recall: 0.96
F1 Score: 0.96
Confusion Matrix: [[11  1]
 [ 1 32]]

Classification Report:              precision    recall  f1-score   support

  Commercial       0.92      0.92      0.92        12
 Residential       0.97      0.97      0.97        33

    accuracy                           0.96        45
   macro avg       0.94      0.94      0.94        45
weighted avg       0.96      0.96      0.96        45
```

KNN- One Neighbor for Data Set



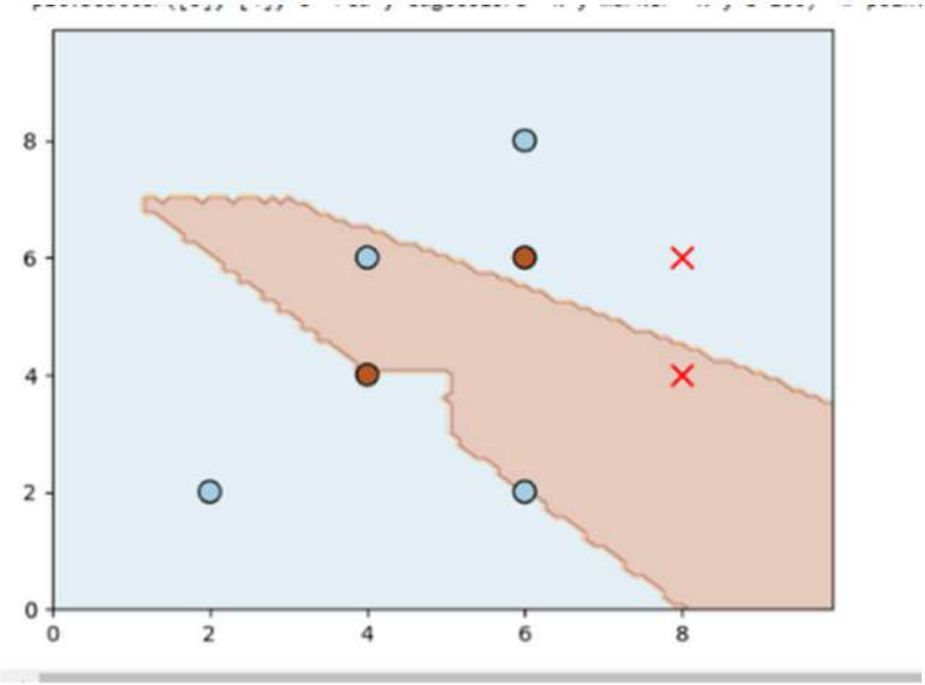
KNN Diagram with One Neighbor

As it is understandable the value for (8,4) remains right in the border of the highlighted area versus the point (8,6) inside of the selected area where (8,4) stands for Blue and (8,6) for yellow figures inside of the studied KNN depending on if switching from 1 to 3 KNN selected in grouped area with predominant values.

Disclaimer: This Booklet only shows answers as a brief document.
For further explanation go through the main document attached.

Results and Investigation

Decision Boundary for K=3: In the Decision Boundary for K=3 the new area covered by the K-Nearest Neighbors which is the following:



KNN with Three Neighbors

```
Accuracy: 0.91
Precision: 0.92
Recall: 0.91
F1 Score: 0.90
Confusion Matrix: [[ 8  4]
 [ 0 33]]

Classification Report:              precision    recall  f1-score   support

  Commercial       1.00      0.67      0.80        12
 Residential       0.89      1.00      0.94        33

    accuracy                           0.91        45
   macro avg       0.95      0.83      0.87        45
weighted avg       0.92      0.91      0.90        45
```

KNN Predictions for Three Neighbors

	Prediction
point (8, 6)	**Blue Square** **Yellow Circle**
point (8, 4)	**Blue Square** **Yellow Circle**

KNN Predictions

Disclaimer: This Booklet only shows answers as a brief document.
For further explanation go through the main document attached.

Results and Investigation

Decision Boundary for K=3: In the Decision Boundary for K=3 the new area covered by the K-Nearest Neighbors which is the following:

	Prediction
point (8, 6)	○ Blue Square ● Yellow Circle
point (8, 4)	● Blue Square ○ Yellow Circle

KNN Predictions

Disclaimer: This Booklet only shows answers as a brief document. For further explanation go through the main document attached.

Results and Investigation

Question 3: Secondary K-Nearest Neighbor Analysis done through modeling.

Predicted labels (+ or -):

Predictions for new points: ['+' '-']

Point [7.81 5.33] is classified as +

Point [9.43 5.29] is classified as -

Point 1: [7.81,5.33] +

Point 2: [9.43,5.29] -

Index of closest neighbors: Indices of the closest neighbors for each point: [[13 11 14]

[11 13 19]]

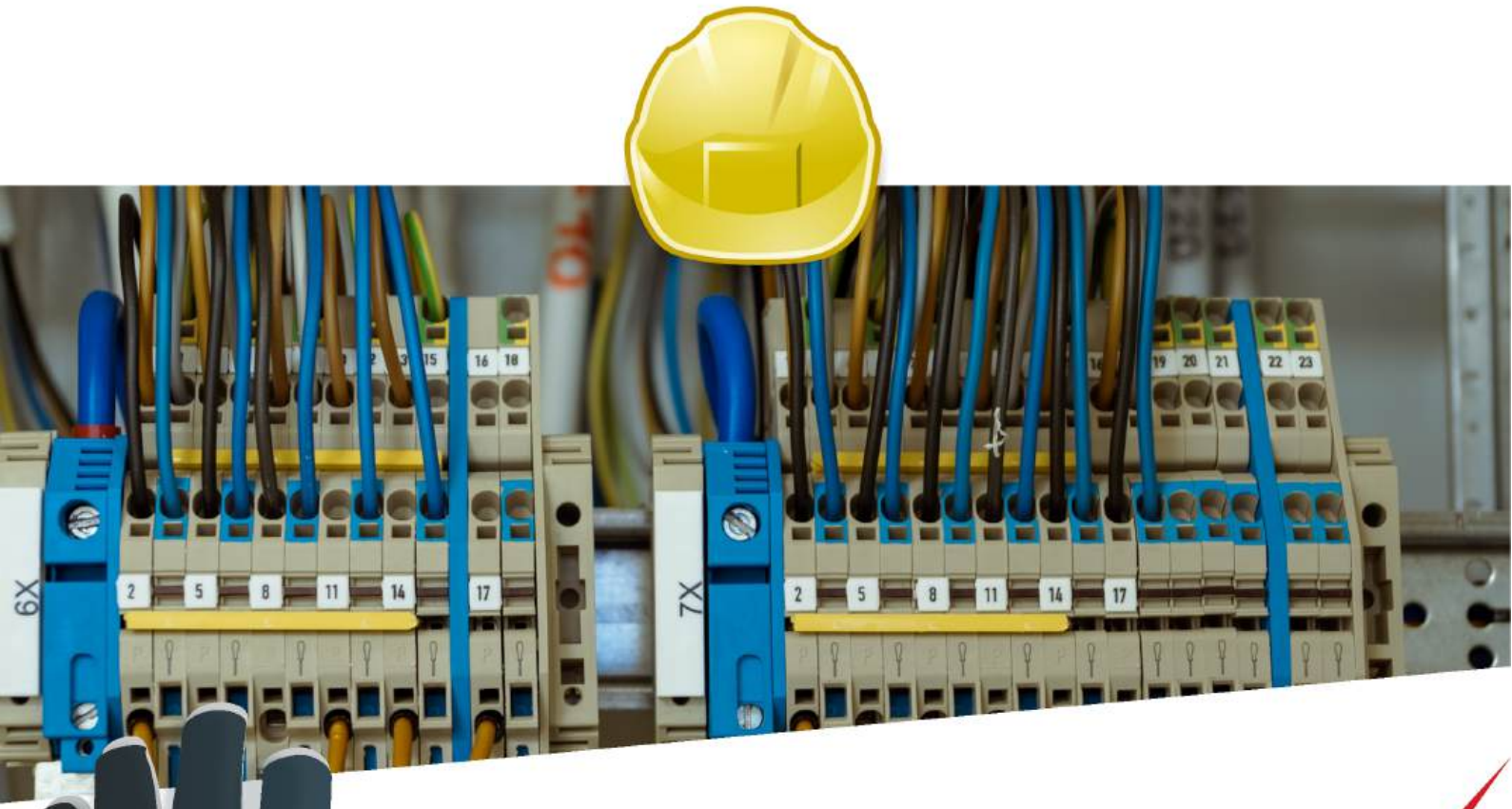
For Point 1: [13,11,14]

For Point 2: [11,13,19]

Disclaimer: This Booklet only shows answers as a brief document.
For further explanation go through the main document attached.

Conclusion

Overall, the importance of keeping any of the subsets and datasets covered with Linear Regression (LR) and the concept of K-Nearest Neighbor is for the variable studies while discovering a novelty into the possibilities regarding various scenarios before dealing with issues and any deductions coming from the current data. Investing in the future for more features to be analyzed like investment can save time and efforts for the business to obtain closer results to the deletion of human error with proper measurements taken into consideration when making a choice or consideration for any commercial or residential client.



Disclaimer: This Booklet only shows answers as a brief document. For further explanation go through the main document attached.

References

Chat Gpt (2024). Chat Gpt. https://chatgpt.com/

Gholibeigian, S. (2024). Classification - KNN.ipynb. https://colab.research.google.com/drive/1eOOh4Kc09rFg9D7vjzPLwyGuHKFBE9gM#scrollTo=cK9x9SfLBfYp

Paz Callizo, W.A. (2024). Construction Company Assignment I Individual ML UCW.ipynb. https://colab.research.google.com/drive/1pilXuoNFniC-XrCb5ZQLCbTxZavr5CnV#scrollTo=zug_npRJoyHi

Disclaimer: This Booklet only shows answers as a brief document.
For further explanation go through the main document attached.