1. Answer: Task description (including task 1, 2 and 3)

Task 1: I tried integrating machine learning research with air pollution epidemiology to support our environment. Precisely, the pollutant 'carbon monoxide (CO)' is being monitored and measured by the 'Halifax' air quality station. I investigated the hourly data on ambient carbon monoxide (CO) in Nova Scotia [1]. Carbon monoxide (CO) is a very poisonous gas resulting from incomplete fuel combustion. Simultaneously, I looked at whether traffic flow impacts roadside carbon monoxide levels and air quality by looking at a particular traffic flow dataset in Nova Scotia [2]. To understand their format, I checked what type of data they contain. The data type of each attribute String may need to be converted to floating values or integers to represent categorical or ordinal values.

Task 2: In the pre-processing steps, as I can see in the CO Dataset Unit, pollutant, and station are not helpful because all have mostly single values, so we will drop them and also convert date and time information into a suitable format and then calculate daily average carbon monoxide (CO) levels specifically for the year 2019. Then, I figured Min-Max scaling to normalize carbon monoxide (CO) levels and discretized them into 'High' and 'Low' based on a threshold of 0.5, creating class labels. After that, the 'Date' column in the 'TRAFFIC_data' DataFrame is converted to DateTime format and sorted in ascending order. This step ensures that the date values are correctly formatted and arranged. Subsequently, the code merges the 'TRAFFIC_data' and 'daily_avg_co_2019' DataFrames, linking them based on the 'Date' column. This merge operation combines the traffic-related data with the daily average carbon monoxide (CO) levels for 2019, allowing for a comprehensive analysis of their relationship and impact. In the final phase of the analysis, I undertook a comprehensive descriptive examination of the dataset consisting of 586 data points. Key attributes such as 'HIGHWAY,' 'SECTION,' 'SECTION LENGTH,' 'ADT,' and 'AADT' displayed varying statistics. The primary focus was on the 'CO_Level,' which exhibited a mean level of 0.112, signifying generally low carbon monoxide (CO) levels. Additionally, I include a summary visualization of data, where I can see that there is a good amount of correlation between ADT and AADT.

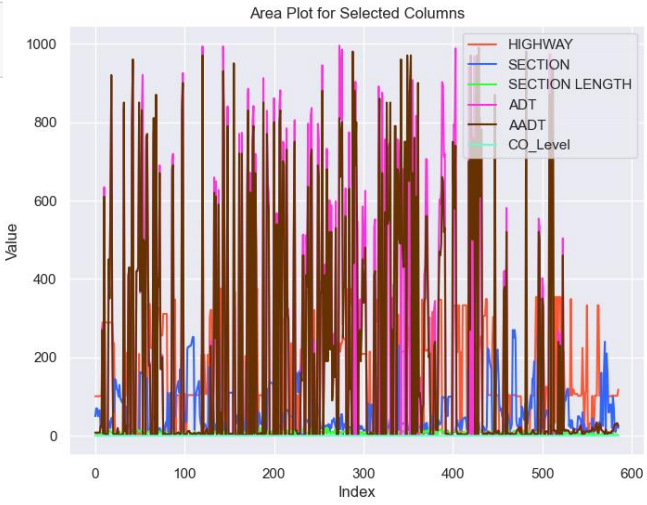
Descriptive Statistics of the datasets

```
In [392]: #a descriptive analysis
          description = merged_dataset.describe()
          print(description)
```

	HIGHWAY	SECTION	SECTION LENGTH	ADT	AADT	\
count	586.000000	586.000000	586.000000	586.000000	586.00000	
mean	148.576792	52.779863	7.401891	197.050370	212.73780	
std	125.552938	56.718809	3.981491	304.282428	303.89292	
min	1.000000	1.000000	0.200000	1.004000	1.00000	
25%	7.000000	17.000000	4.150000	2.838750	3.60000	
50%	104.000000	30.000000	7.253500	8.162000	10.95000	
75%	245.000000	60.000000	10.040000	381.500000	430.00000	
max	376.000000	270.000000	20.720000	995.000000	990.00000	

	CO_Level	CO_Level_normalized
count	586.000000	586.000000
mean	0.112084	0.396045
std	0.024675	0.146586
min	0.045417	0.000000
25%	0.092917	0.282178
50%	0.110833	0.388614
75%	0.124583	0.470297
max	0.186875	0.840347

Visualization of data



Task 3: I've chosen to work with the following features: 'HIGHWAY,' 'SECTION,' 'SECTION LENGTH,' 'ADT,' and 'AADT' (Average Daily Traffic). These features have been selected because they are expected to have a significant impact on carbon monoxide (CO) levels in the Halifax city area.

- HIGHWAY: This feature represents the type of highway, and it's important because different types of highways can have varying traffic patterns, road conditions, and, in turn, different levels of CO emissions.
- SECTION: The 'SECTION' feature is crucial as different sections of a highway may have distinct characteristics that influence CO levels.
- SECTION LENGTH: The length of a highway section is relevant because shorter sections may have different traffic and environmental conditions compared to longer sections, potentially affecting CO levels.
- ADT (Average Daily Traffic): 'ADT' is a vital indicator of daily traffic volume, and it's included because higher traffic volumes often lead to increased CO emissions.
- AADT (Annual Average Daily Traffic): 'AADT' provides insights into long-term traffic trends, helping to account for seasonal variations that can impact CO levels.

The target variable chosen for this task is 'CO_Level_discretized.' It's a binary classification variable with two classes: 'High' when CO levels are greater than or equal to 0.5 and 'Low' when CO levels are less than 0.5. This choice aligns with the goal of classifying CO levels as 'High' or 'Low,' as outlined in the problem statement.

These feature selections have been made to comprehensively capture factors that could influence CO levels in Halifax, taking into account different types of roads, specific road sections, their lengths, and traffic volumes. The target variable choice corresponds with the specified classification objective.

A classification task [3] is a machine learning task where the goal is to predict the category of a new data point. Suppose, in this above problem, the classification task involves categorizing CO_levels in Halifax as 'High' or 'Low'.

A decision tree [4] is a machine learning model that can be used for both classification and regression tasks. It is a tree-like structure where each node represents a decision, and each leaf node represents a class. A decision tree is a reasonable model to try for the classification task in this case because:

- The data is structured. Each data point has a set of features, and the goal is to predict the category of the data point based on those features.
- The dataset is likely to contain a combination of categorical features (e.g., 'HIGHWAY,' 'SECTION') and numerical features (e.g., 'ADT' and 'AADT'). Decision Trees can naturally handle both types, making them suitable for this dataset.
- There are a relatively small number of classes. This means that the decision tree will not be too complex and will be able to learn the relationships between the features and the classes.

2. Answer: question i (Task 4)

The most influential factor for CO level can be determined by calculating the Information Gain (IG) for each attribute when used as a split criterion in a decision tree. IG measures the reduction in uncertainty (entropy) that an attribute provides when used for splitting the data. The attribute with the highest IG is considered the most influential factor because it contributes the most to reducing the uncertainty in predicting the CO level. In our analysis, the attribute with the highest IG is ADT, with an IG of 0.6442. This indicates that ADT is the most influential factor for CO levels. The high IG of ADT suggests that it provides the most useful information for classifying CO levels. This might be because ADT is closely related to traffic flow, and higher traffic volumes can lead to increased CO levels due to vehicle emissions.

3. Answer: question ii (Task 4)

Use the Information Gain (IG) as the decision criterion [5] to select which attribute to split on. Below showing the calculations procedures for the IG for the root node.
The Information Gain (IG) is calculated as follows:

- Calculate the entropy of the parent node (entropy_parent).
- For each attribute, calculate the weighted average of the entropy of child nodes (weighted by the number of instances in each child node) and subtract this from the entropy_parent.
- Select the attribute with the highest IG as the root node for splitting.

Here's the IG calculation for the root node:

- Calculate the entropy of the parent node (entropy_parent):
  entropy_parent = -p_high * log2(p_high) - p_low * log2(p_low)
  Where p_high is the proportion of instances with CO level ≥ 0.5, and p_low is the proportion of instances with CO level < 0.5.
- Calculate the IG for each attribute:
  IG_attribute = entropy_parent - weighted_entropy_child
  Where weighted_entropy_child is the weighted average of entropy in the child nodes after splitting on the attribute.
  Repeat this calculation for all attributes and select the one with the highest IG as the root node for splitting.

Here’s the information gain output from using this procedure:

```
Information gain for:HIGHWAY -- 0.2224828669132049
Information gain for:SECTION -- 0.16971134646773953
Information gain for:SECTION LENGTH -- 0.536753356780813
Information gain for:ADT -- 0.6441521570701433
Information gain for:AADT -- 0.4758156333259556
Attribute with the highest Information Gain (most_influential_factor): ADT
Information Gain for the root node: 0.6441521570701433
```

4. Answer: question iii (Task 4)

(a) Fit a decision tree with the default parameters on 50% of the data and test it on 50% held-out data. Employing a decision tree classifier, the model achieved an accuracy of approximately 81.23%. It demonstrated stronger performance in identifying the "Low" class while encountering difficulties with the "High" class.
The classification report provided the following insights:

```
Confusion Matrix:
 [[ 22  24]
 [ 31 216]]
Classification Report:
              precision    recall  f1-score   support

        High       0.42      0.48      0.44        46
         Low       0.90      0.87      0.89       247

    accuracy                           0.81       293
   macro avg       0.66      0.68      0.67       293
weighted avg       0.82      0.81      0.82       293

Accuracy: 0.8122866894197952
Number of mislabeled points out of a total 293 test points : 55
```

The report demonstrates that the model's precision, recall, and F1-score varied between the "High" and "Low" classes. Specifically, it exhibited a higher F1-score for "Low" and a lower F1-score for "High." Moreover, the model achieved an accuracy of 81.23%, but it misclassified 55 out of 293 test points, indicating potential areas for refinement.

(b) I applied a 10-fold cross-validation technique to build another model for predicting CO levels. The model achieved an accuracy of approximately 86.18% and demonstrated consistent performance across the folds. The results are presented in the classification report as follows:

```
Confusion Matrix:
 [[ 65  40]
 [ 41 440]]
Classification Report:
              precision    recall  f1-score   support

        High       0.61      0.62      0.62       105
         Low       0.92      0.91      0.92       481

    accuracy                           0.86       586
   macro avg       0.76      0.77      0.77       586
weighted avg       0.86      0.86      0.86       586

Accuracy for each fold: [0.88135593 0.83050847 0.89830508 0.86440678 0.77966102 0.89830508
 0.87931034 0.84482759 0.89655172 0.84482759]
Mean Accuracy: 0.8618059614260666
```

The classification report reveals that the model's performance, assessed through precision, recall, and F1-score, differs between the "High" and "Low" classes. It achieves a higher F1-score for "Low" (0.92) than for "High" (0.62), indicating a better ability to predict "Low" instances. The model achieves an overall accuracy of 86.18%, with a macro average F1-score of 0.77. Moreover, I computed the accuracy for each fold in the 10-fold cross-validation, with values ranging from 0.78 to 0.90. The mean accuracy across all folds is approximately 0.86, suggesting a consistent and reliable model performance.

(c) The model makes sense for several reasons evident in the evaluation results. In the 10-Fold Cross-Validation, the model consistently achieved higher precision, recall, and F1-scores for both the "High" and "Low" classes. This suggests that the model reliably distinguishes between classes and is capable of maintaining consistent performance across multiple data subsets. Moreover, the mean accuracy of 86.18% in the 10-Fold Cross-Validation demonstrates the model's robustness and ability to generalize its performance effectively. It consistently outperforms the Hold-Out evaluation, where the accuracy was 81.23%.

In our decision tree model, the minimum leaf depth is observed to be 3, which suggests that the shallowest leaf nodes are at a depth of 3 levels from the root. A minimum depth of 3 is generally reasonable and indicates that the leaf nodes are not extremely small or shallow. Each leaf node at this depth likely contains a reasonable number of instances, which is a positive sign for model generalization.

(d)

```
Confusion Matrix:
 [[ 22  24]
 [ 31 216]]
Classification Report:
              precision    recall  f1-score   support

        High       0.42      0.48      0.44        46
         Low       0.90      0.87      0.89       247
```

Hold-Out evaluation matric

- **True Positives (TP)**: 216
- **True Negatives (TN)**: 22
- **False Positives (FP)**: 24
- **False Negatives (FN)**: 31

```
Confusion Matrix:
 [[ 65  40]
 [ 41 440]]
Classification Report:
              precision    recall  f1-score   support

        High       0.61      0.62      0.62       105
         Low       0.92      0.91      0.92       481
```

10-Fold Cross-Validation evaluation matric

- **True Positives (TP)**: 440
- **True Negatives (TN)**: 65
- **False Positives (FP)**: 40
- **False Negatives (FN)**: 41

Based on the provided evaluation matrices [6] for both Hold-Out and 10-Fold Cross-Validation, I believe that 10-fold cross-validation is the more suitable approach for my data and model. This preference is based on the following reasons:

- In the confusion matrix for the Hold-Out evaluation, the model correctly predicted 22 instances in the 'High' category and 216 instances in the 'Low' category. However, it also made some incorrect predictions. In contrast, in the confusion matrix for the 10-Fold Cross-Validation, the model correctly predicted 65 'High' instances and 440 'Low' instances across multiple folds.
- In the Hold-Out evaluation, the model displayed precision, recall, and F1-scores of 0.42, 0.48, and 0.44 for the "High" class, and 0.90, 0.87, and 0.89 for the "Low" class, respectively. However, in the 10-Fold Cross-Validation, we observed higher values across the board. The "High" class in the 10-Fold Cross-Validation showed a precision of 0.61, recall of 0.62, and an F1-score of 0.62. For the "Low" class, precision is 0.92, recall is 0.91, and F1-score is 0.92. These values suggest that the model performed consistently better in 10-Fold Cross-Validation, with higher precision and recall for both "High" and "Low" classes.

This suggests that the model is more reliable and better at adapting to different data scenarios. Therefore, 10-Fold Cross-Validation is the better choice for evaluating this data and model.

5. Summary of my results:

In summary, the 10-Fold Cross-Validation method consistently outperformed the Hold-Out evaluation, demonstrating higher precision, recall, and F1-scores for both "High" and "Low" classes. While Hold-Out yielded an accuracy of 81.23%, the 10-Fold Cross-Validation showed superior performance with a mean accuracy of 86.18%. This indicates that the model is not only more robust but also better at generalizing its performance across different data subsets. Therefore, 10-Fold Cross-Validation is the preferred choice for evaluating this data and model.

6. References:

[1] https://data.novascotia.ca/Environment-and-Energy/Nova-Scotia-Provincial-Ambient-Carbon-Monoxide-CO-/8tvc-9ah2
[2] https://data.novascotia.ca/Roads-Driving-and-Transport/Traffic-Volumes-Provincial-Highway-System/8524-ec3n
[3] https://www.sciencedirect.com/topics/computer-science/classification-task
[4] https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/cem.873
[5] https://towardsdatascience.com/entropy-and-information-gain-in-decision-trees-c7db67a3a293
[6] https://www.geeksforgeeks.org/metrics-for-machine-learning-model/