Academic Year	Module	Assessment Number	Assessment Type
2024	5CS037-Concept and Technologies of AI	Final	Report

CLASSIFICATION TASK

Student Id : 2408620

Student Name : Rodik Awal

Section : L5CG7

Module Leader : Siman Giri

Tutor : Siman Giri

Submitted on : 11/02/2025

Contents

Classification Report

Introduction

This report is based on a classification task performed on a dataset of CO2 emissions, which is also related to SDG 13 – Climate Action. Model building, Exploratory Data Analysis(EDA), Data Preprocessing, and evaluation with two separate machine learning classifiers are all parts of the task.

Dataset

The “CO2 Emissions.csv” dataset, which includes a numbers of features which relate to vehicle specifications and their corresponding CO2 emissions, is well structured to predict CO2 emission levels and locate factors that contribute.

Objective

The aim of this model is to develop a classification model that offers insights into the factors influencing emissions and accurately predicts the target variable (CO2 emission).

Exploratory Data Analysis (EDA) Statistical Implementation and Visualization

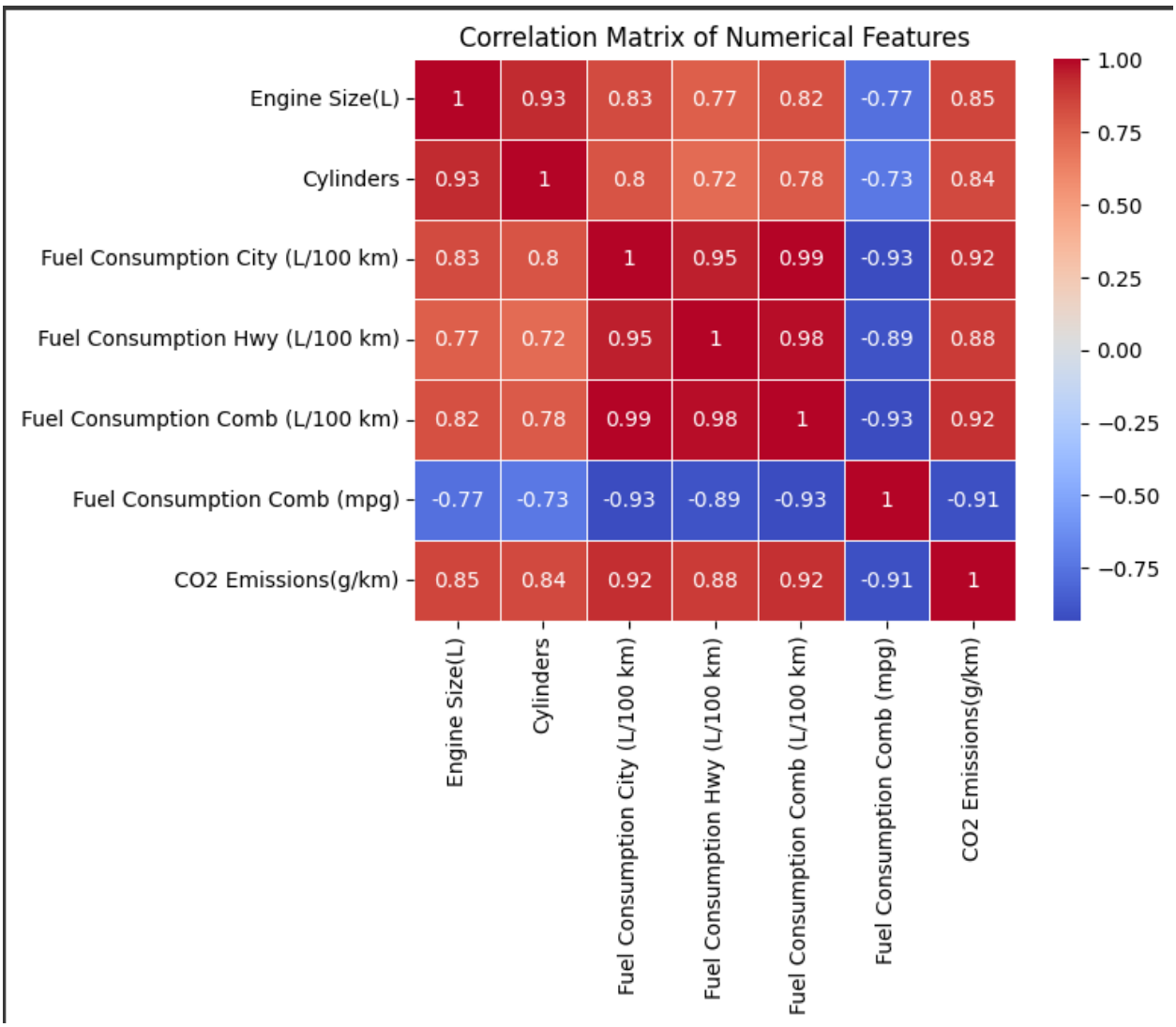
1. The dataset was reviewed for any missing values before summary statistic were generated.

	Engine Size(L)	Cylinders	Fuel Consumption City (L/100 km)	Fuel Consumption Hwy (L/100 km)	Fuel Consumption Comb (L/100 km)	Fuel Consumption Comb (mpg)	CO2 Emissions(g/km)
count	5956.00	5956.00	5956.00	5956.00	5956.00	5956.00	5956.00
mean	3.17	5.63	12.65	9.09	11.05	27.33	251.63
std	1.37	1.85	3.56	2.29	2.96	7.18	59.27
min	0.90	3.00	4.20	4.00	4.10	11.00	96.00
25%	2.00	4.00	10.10	7.50	8.90	22.00	208.00
50%	3.00	6.00	12.10	8.70	10.60	27.00	246.00
75%	3.80	6.00	14.70	10.30	12.70	32.00	290.00
max	8.40	16.00	30.60	20.60	26.10	69.00	522.00

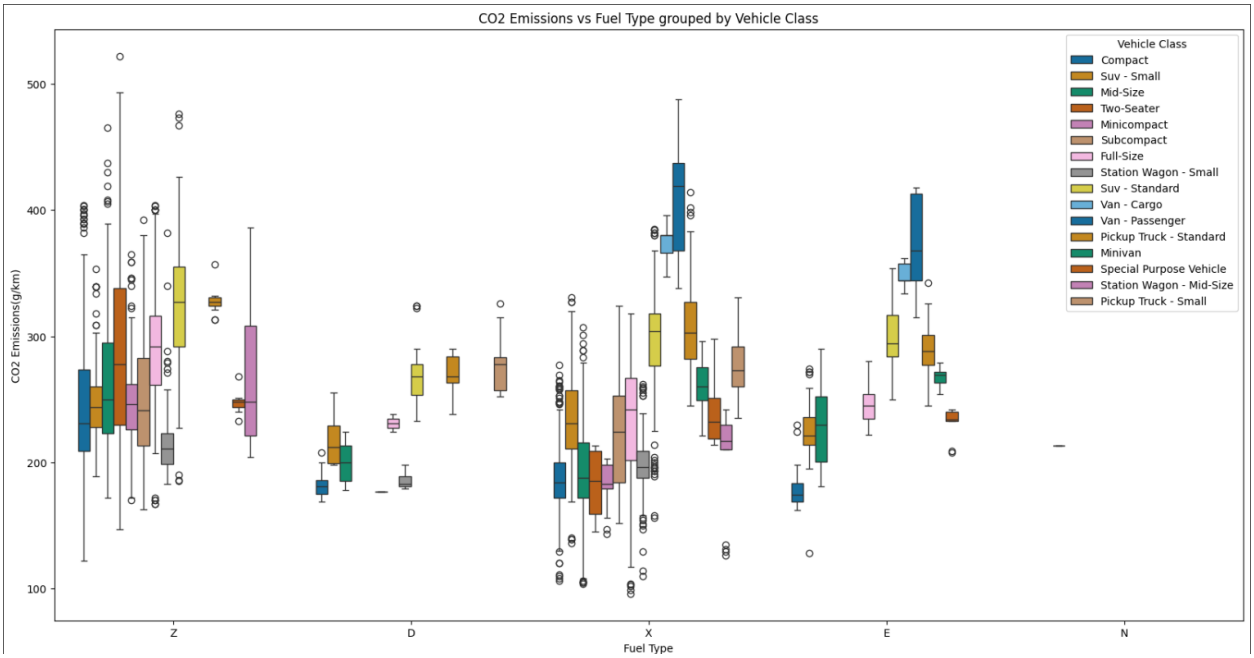
	0
Make	0
Model	0
Vehicle Class	0
Engine Size(L)	0
Cylinders	0
Transmission	0
Fuel Type	0
Fuel Consumption City (L/100 km)	0
Fuel Consumption Hwy (L/100 km)	0
Fuel Consumption Comb (L/100 km)	0
Fuel Consumption Comb (mpg)	0
CO2 Emissions(g/km)	0

dtype: int64

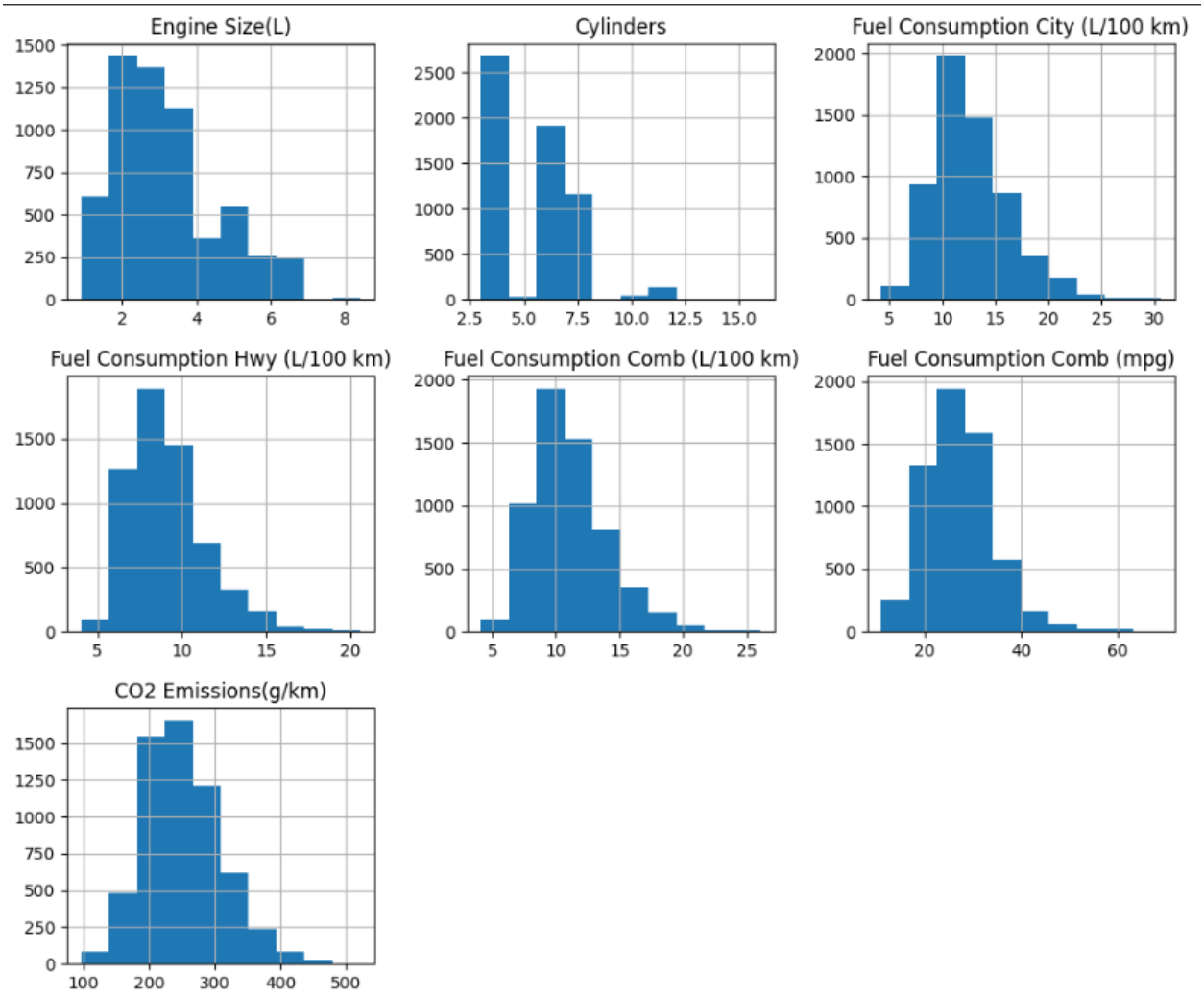
2. Data was transformed into numerical features, and a correlation matrix was created.



3. Boxplot of CO2 Emissions vs Fuel Type grouped by Vehicle Class was created to find out these features relate to one another and look for errors.



4. To see the distribution and connections between different features and how they affect CO2 Emissions, histograms of every feature were created.

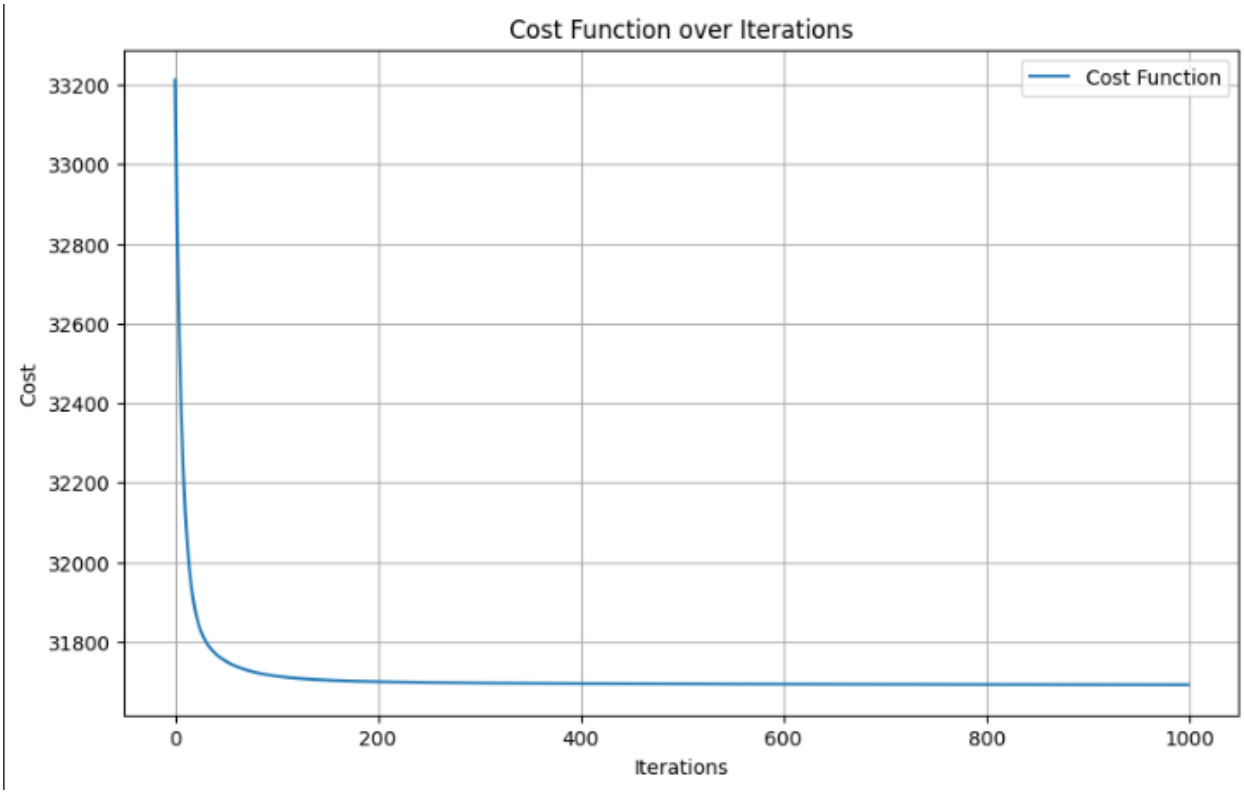


Building model from Scratch

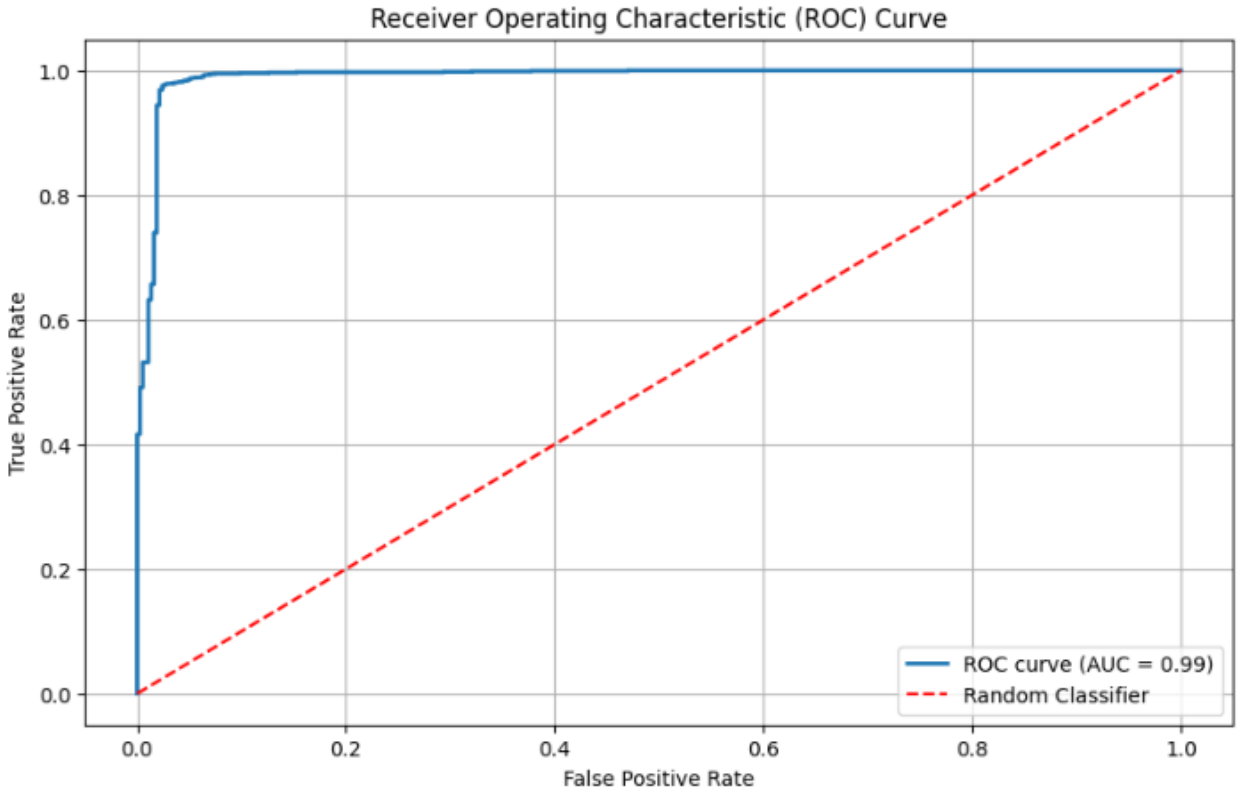
- A logistic regression model that was created from the scratch had an accuracy of 0.0772

```
Accuracy: 0.0772
```

- As the cost drops over iterations, the cost function graph helps evaluate learning progress and illustrates how well the model is learning.



- The True Positive Rate(TPR) against False Positive Rate (FPR) tradeoff at various levels is shown by the ROC curve. Overall performance is indicated by the Area Under Curve(AUC) with a value near to 1 showing good performance.
- The figure shows AUC of 0.99 which is close to 1 indicating good performance.



Evaluation of the Model

The best performing model was identified by classifying the differences between the models using the logistic regression and random forest models.

- Logistic regression model (Model 1) has the accuracy of 0.9732
- Random regression model (Model 2) has the accuracy of 0.9790

```
Model 1 (Logistic Regression) Accuracy: 0.9732
Model 2 (Random Forest) Accuracy: 0.9790

Logistic Regression Report:
              precision    recall  f1-score   support

           0       0.92      0.95      0.94       243
           1       0.99      0.98      0.98       949

    accuracy                           0.97      1192
   macro avg       0.95      0.96      0.96      1192
weighted avg       0.97      0.97      0.97      1192


Random Forest Report:
              precision    recall  f1-score   support

           0       0.95      0.95      0.95       243
           1       0.99      0.99      0.99       949

    accuracy                           0.98      1192
   macro avg       0.97      0.97      0.97      1192
weighted avg       0.98      0.98      0.98      1192
```

Hyper-Parameter

```
Best Parameters for Logistic Regression: {'C': 100, 'solver': 'liblinear'}
Optimized Logistic Regression Accuracy: 0.9740
```

Following hyper-parameter adjustment, the optimum accuracy of logistic regression is 0.9740, indicating no change in accuracy.

```
Best Parameters for Random Forest: {'max_depth': None, 'min_samples_split': 10, 'n_estimators': 200}
Optimized Random Forest Accuracy: 0.9765
```

The optimum accuracy of random forest after hyper-parameter tuning is 0.8696, indicating that the model's performance has fallen but still surpasses logistic regression.

REF (Recursive Feature Elimination)

REF is imported in order to choose the best features, which will enhance model performance by choosing the most relevant features and recursively removing the least significant ones.

```
Selected Features for Logistic Regression: ['Engine Size(L)', 'Cylinders', 'Fuel Consumption City (L/100 km)', 'Fuel Consumption Hwy (L/100 km)']
Selected Features for Random Forest: ['Engine Size(L)', 'Cylinders', 'Fuel Consumption City (L/100 km)', 'Fuel Consumption Hwy (L/100 km)']
```

Results

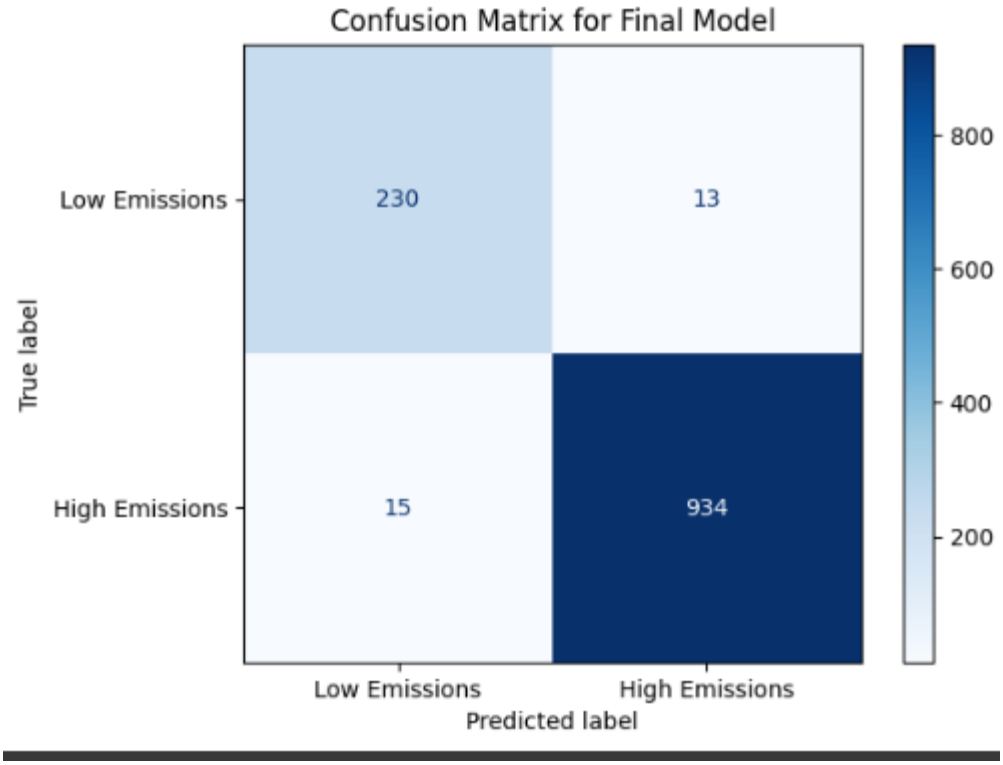
Logistic regression performed slightly better than Random Forest regression.
The figure below shows the final model ability to distinguish between Low Emission and High Emission.

```
Final Model Accuracy: 0.9765

Final Model Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.95      0.94       243
           1       0.99      0.98      0.99       949

    accuracy                           0.98      1192
   macro avg       0.96      0.97      0.96      1192
weighted avg       0.98      0.98      0.98      1192
```



Conclusion

This report used machine learning techniques to predict CO2 emission levels through a classification task. Random forest and logistic regression models were used and assessed following data preprocessing and exploratory data analysis. After hyper-parameter tuning, logistic regression outperformed random forest in terms of accuracy. Both models successfully differentiated between high and low CO2 emissions, providing insights into the factors influencing emission levels, despite some performance variance. These findings show how AI models can be used to support SDG 13 and other climate action initiatives.