Portfolio: Regression Task Report

Submitted

by:

Sasika Bista

Level 5

Group 5

Submitted to:

Faculty of Concept And Technologies

Of AI

Bachelor Of Computer Science

Herald College Kathmandu

February 10, 2026

Abstract

This study aims to classify countries based on CO_2 emission levels using the "CO_2 emission by countries year wise (1750-2022)" dataset. Preprocessing and exploratory data analysis (EDA) were performed to handle missing and inaccurate values and to prepare the data for modeling. Two classification models were implemented: Logistic Regression and Decision Tree Classifier. Model performance was evaluated using accuracy, precision, recall, and F1-score metrics. Feature selection was performed using the wrapper method with Recursive Feature Elimination (RFE), and hyperparameter tuning was conducted with GridSearchCV. The results show that the Decision Tree classifier outperforms Logistic Regression, effectively capturing non-linear patterns and skewed distributions in the data, highlighting the importance of feature selection and hyperparameter optimization for accurate classification of CO_2 emission levels.

Contents

Table of figures

1. Introduction

1.1. Problem Statement

The goal of this report is to predict Emission_Class variable using classification techniques. This class was indirectly created using the given feature “CO2 emission (Tons)” and diving it into binary values where 1 meant high emission and 0 meant low emission.

1.2. Dataset

The dataset is sourced from Kaggle, an open data platform. It is titled “CO_2 emission by countries year wise (1750-2022)” and is publicly available for research and analysis. This provides comprehensive annual CO_2 emission data for 220 unique countries spanning from 1750 to 2020, along with country-level information such as population, area, density, and the percentage of world landmass covered. This dataset aligns with (United Nations Sustainable Development Goals) UNSDG’s, SDG 13: Climate Action, as it focuses on predicting CO_2 emissions and identifying countries with high emission levels. By analyzing historical emission trends and predicting future values, which can help inform climate policies, enabling governments and organizations to implement effective mitigation strategies indirectly also helping to improve healthy environment.

1.3. Objective

The objective of this analysis is to build a predictive classification model thet estimates the Emission_class of a country based on its features such as population, area, density, and landmass covered by the country in a given year.

2. Methodology

2.1. Data Preprocessing

The data contained mismatched values for necessary features such as population and area. Although, the co2 emission was provided for each year the population and area as well as other features derived from this were provided for the year 2022 only, while the data set also lacked the data for the year 2021 and 2022. So three external data sets were used to fill the missing and mismatched values.

For the missing data for the population feature, **Global CO2 & Greenhouse Gas Emission Year wise (1750-2024)** , for area features, **Land_area_km.csv** from the world bank, and for the missing year 2021 and 2022 data were imported from the **Annual CO2 Emission (1750-2022)** was used.

However, since each datasets had unique country values, some more or less than those in our primary dataset, they were filtered out to include only the countries present in the primary dataset, whereas, countries from the primary datasets that didn't occur in any of the other datasets were dropped entirely.

On further analysis, it was observed that the Co2 emission for the years 2021 and 2022, despite enlisted to be in tons (same unit as in the primary set) were observed to be have ten time lower than values from 2020. Since the overall trend of CO2 Emission from the year 1750 to 2020 was increasing , a ten times difference seemed illogical. This could have been resulted due to unspecified units in these both dataset , it said in tons, but failed to mention if they were megatons, gigatons or just tons. As such the newly inserted rows for year 2021 and 2022 were dropped entirely.

A new feature Co2_pper_capital was created to normalize the CO2 emissions by populations. As total emission alone can be miss leading for the countries with larger population. And this continuous value was converted in to a categorical measure to create a categorical target variable knows as Emission_Class for the binary classification task.

Finally, inter-quartile range was calculated and by using the 1.5 rule outliers where filtered out for the co2 emission in tons. There exist only upper bound outliers in the dataset where it could be observed that for similar values of population and area the co2 emission varies different. The total count of these outliers were 11554. However, since the data set experience a scarcity of the higher emission values upon removing these outliers wipes out nearly half of the high emission indexes. So in order to not have our models experience overfitting , it was decided to keep these outlier values.

1.1. Exploratory Data Analysis (EDA)

EDA tasks were performed via visualization such as bar graph, scatter plot and line chart, basic statistics, and outliers.

Basic statistics revealed that the average co2 emission per capital was around 0.0462, suggesting that the co2 emission per cap is very low across the dataset. It's median was 0.0 which is result of having the dataset from early industrial revolution years , during which most countries had little to no industrial activity and therefore negligible emissions as such the large number of the dataset was 0. Similarly standard deviation of the co2 emission was 0.21, which suggests that the data was highly dispersed and not evenly distributed.

A observation of co2 emission trends shows that the co2 emission has increased by a lot over time, with emission remaining very low and almost flat from 1750 to mid-1800. A gradual rise begin from 1950 and since then a sharp and accelerated increase can be noticed. This highlights the impact of industrialization, population, and fossil fuel consumption growth , post 1950 period which marks rapid expansion in industrial production, energy use, and global economic activity .

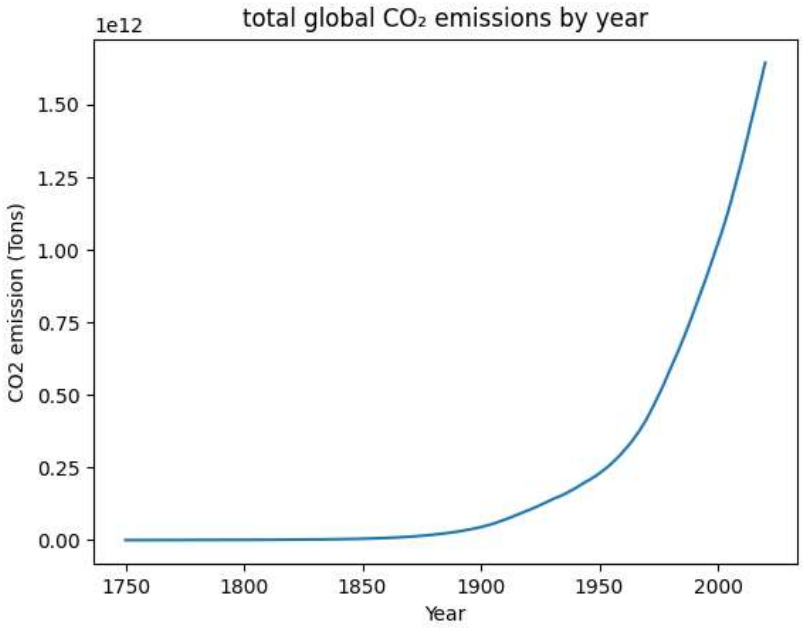

Figure 1Total co2 emission by years

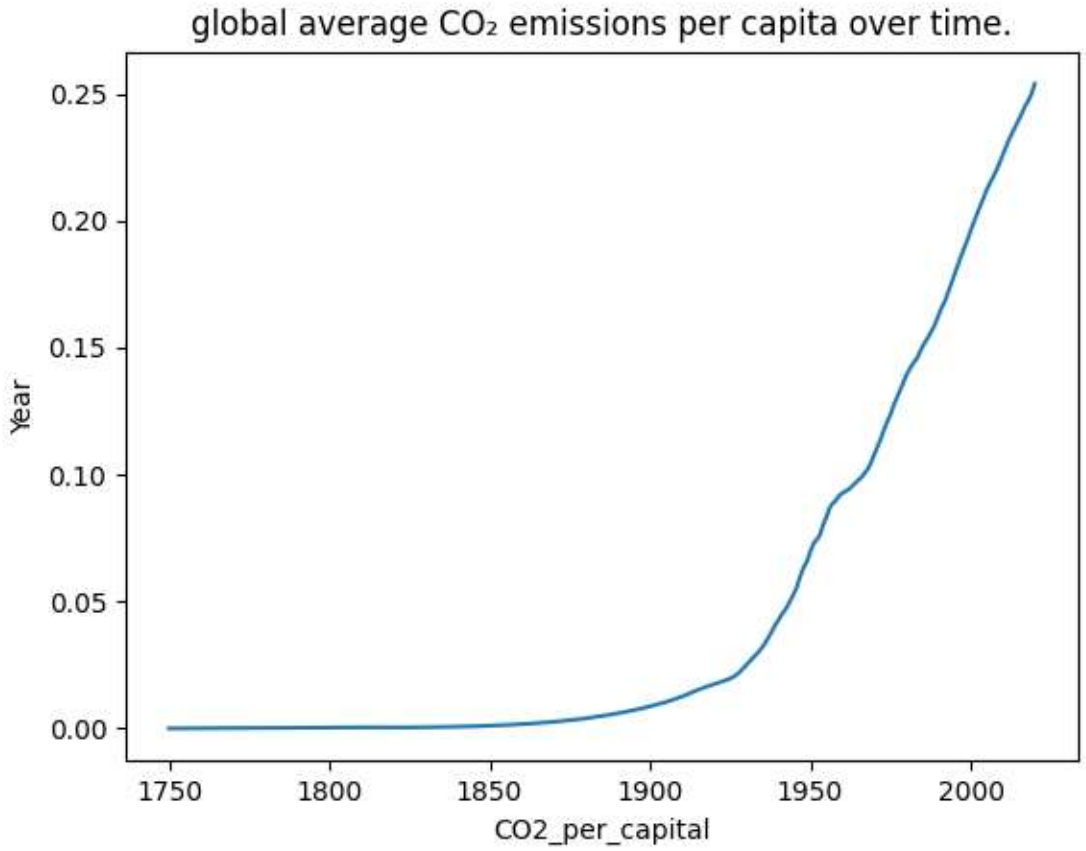

Figure 2 line chart showing the global average CO_2 emissions per capita over time.

Similarly, observation of global average co2 emission similar trends and increase rate of over time.

By selection 5 of the top ten industrialized countries China, UK, USA, Japan, India , can further represent the global emission changes through out years. It can be seen that form an early 1750's UK shows an early rise , reflected as one of the first countries to industrialize during the industrial revolution.

Similarly, USA shows a strong and steady increase , especially after the late 1800s, even becoming a higher per-capital emitters than UK by 2020.

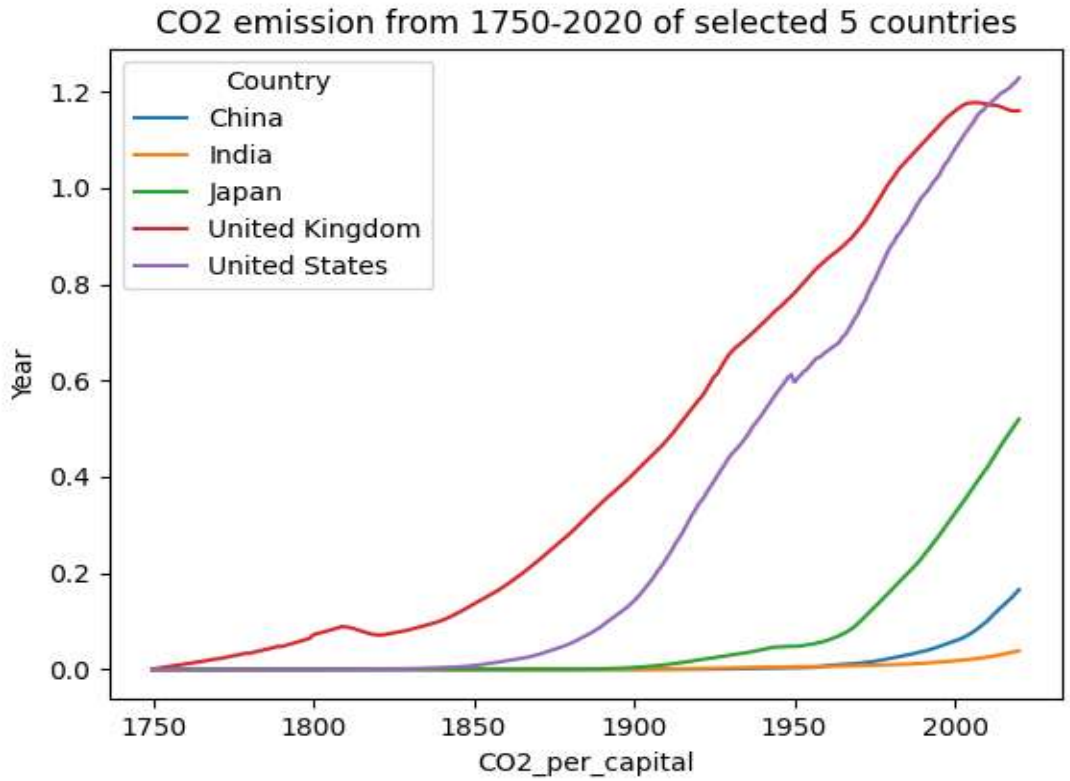

Figure 3line chart showing CO2 emission from 1750-2020 of selected 5 countries.

Japan , China, and India on the other hand were comparatively flat, meaning low emitters up until 1850s from where Japan can be seen taking a slow rise and significated growth after the mid 1800s , mainly after World War II. Whereas China, rise only after 2000 reflecting recent developments. And finally, India still rising gradually , reflecting slower development compared to other developed countries.

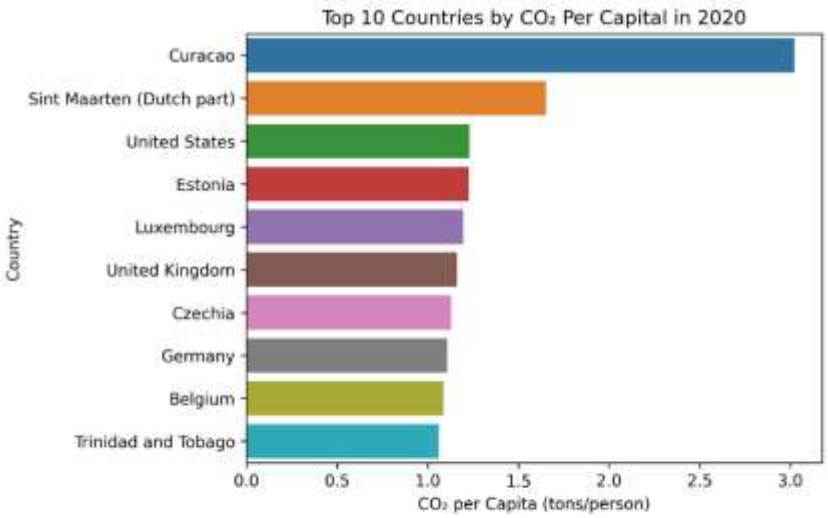

Figure:4 Bar chart showing top 10 Countries by CO2 per capital in 2020

Figure 4 Bar chart showing top 10 Countries by CO2 per capital in 2020.

Further analysis on the top 10 co2 emission countries for the year 2020 reveals that even developing countries with energy-intensive industries, tourism and energy use, and oil an d gas production can lead to higher per capital emission than some most developed countries. United States also ranges highest among the developed countries, while many European countries still have relatively high per capital emission.

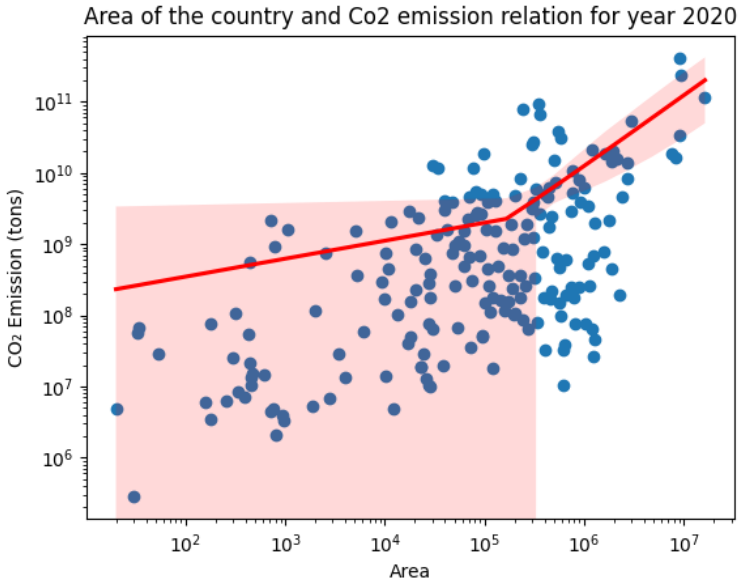

Figure 5 Scatter plot showing the relation between Area and the CO2 emission of a country for the year 2020

In the visualization of CO2 emission and Area of a country, the red line indicates that the relation between larger areas and the co2 has a positive trend. However, the shaded region highlights the lower-age range, where emissions vary substantially despite relatively small difference in land size. This suggest Area alone cannot determine the CO2 emission of a country.

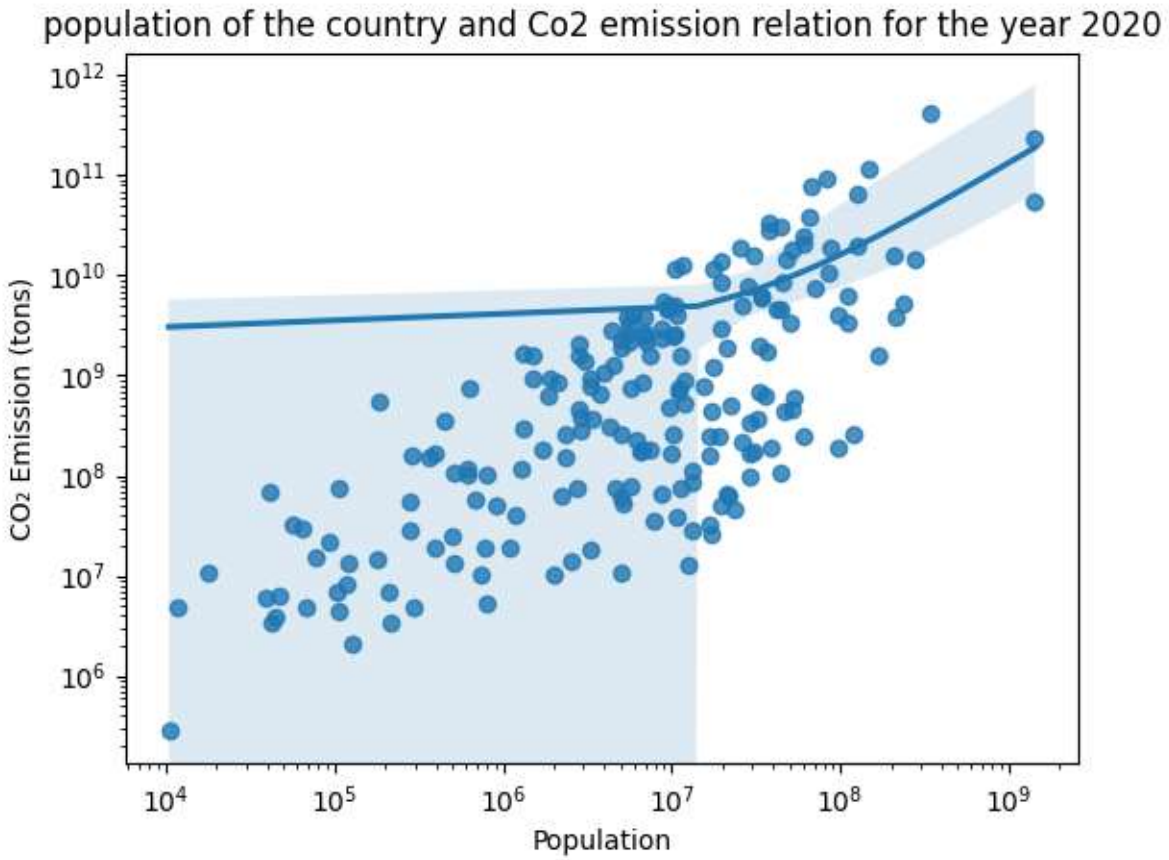

Figure 6 Scatter plot showing the relation between population and the CO2 emission of a country for the year 2020

The scatter plot between the population and the CO2 emission, shows a positive relationship between population size and total co2 emission. As population increases, emission tend to rise. However, there is still noticeable variation among countries with similar population size, suggesting that the population also cannot become a sole factor in determining the co2 emission.

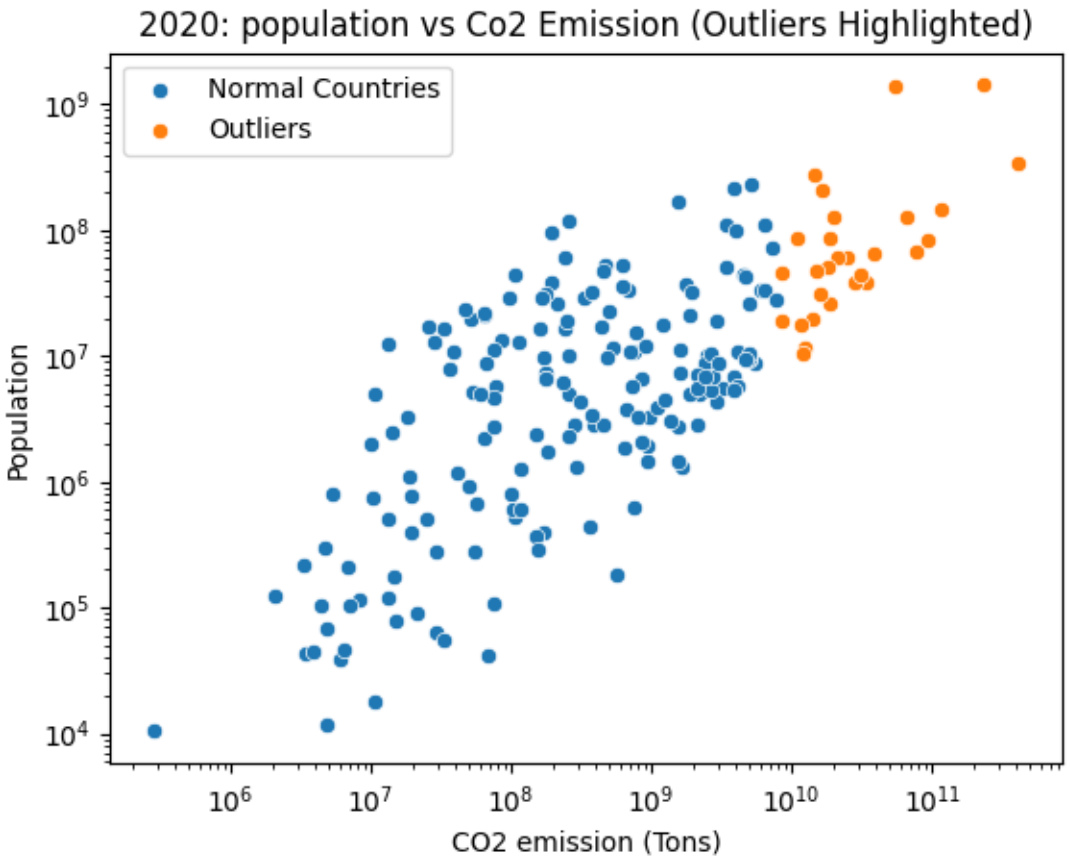

Figure 7 Scatter plot showing normal points and outliers\n among population vs Co2 Emission for year 2020.

Outlier detection of co2 emission for the year 2020, further highlights how some countries emits much more CO2 than expected for their population. Normal points can be seen to follow the trend , high population high emission where as the upper-bound outliers indicating high population and even higher emission that suggest these countries emit way more emission than those countries of same population range.

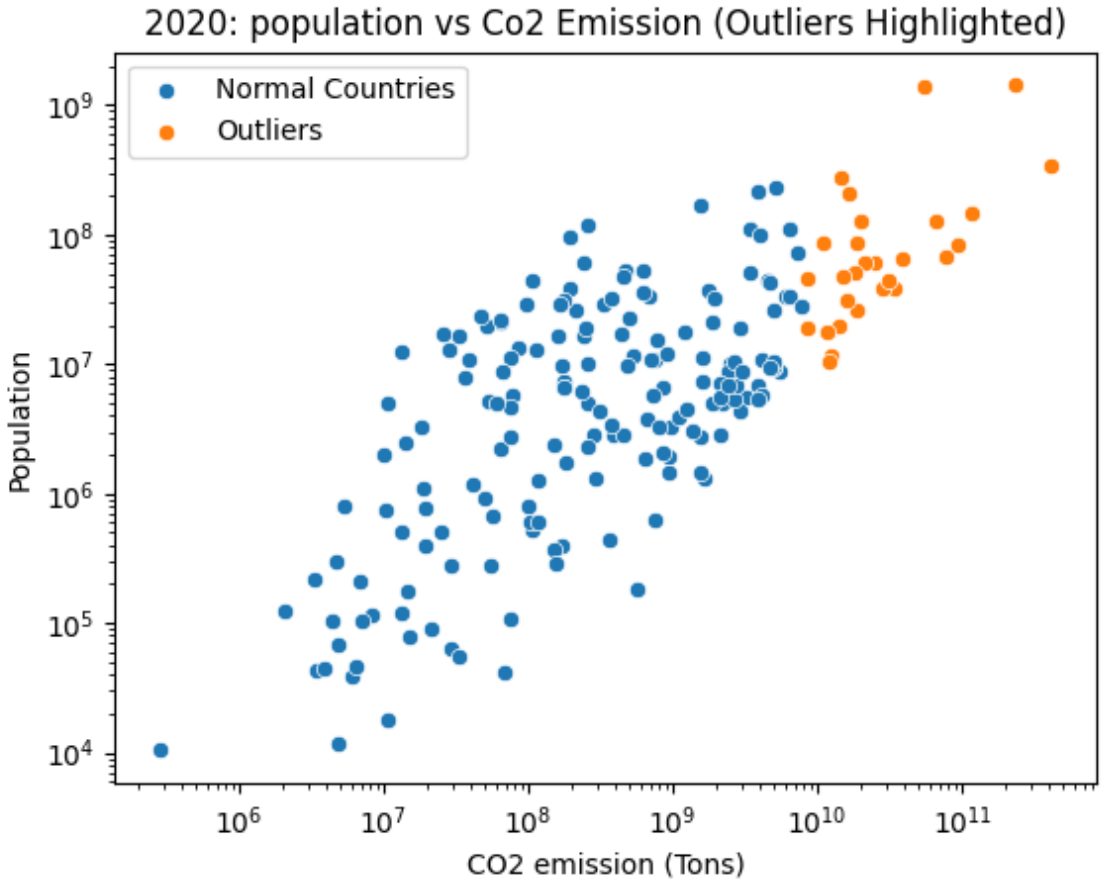

Figure 8 Scatter plot showing normal points and outliers\n among population vs Co2 Emission for year 2020.

Similar behavior can be observed in terms of population as well. Where , as population increases co2 emission also increases however few countries emit higher co2 emission than those countries with similar population values.

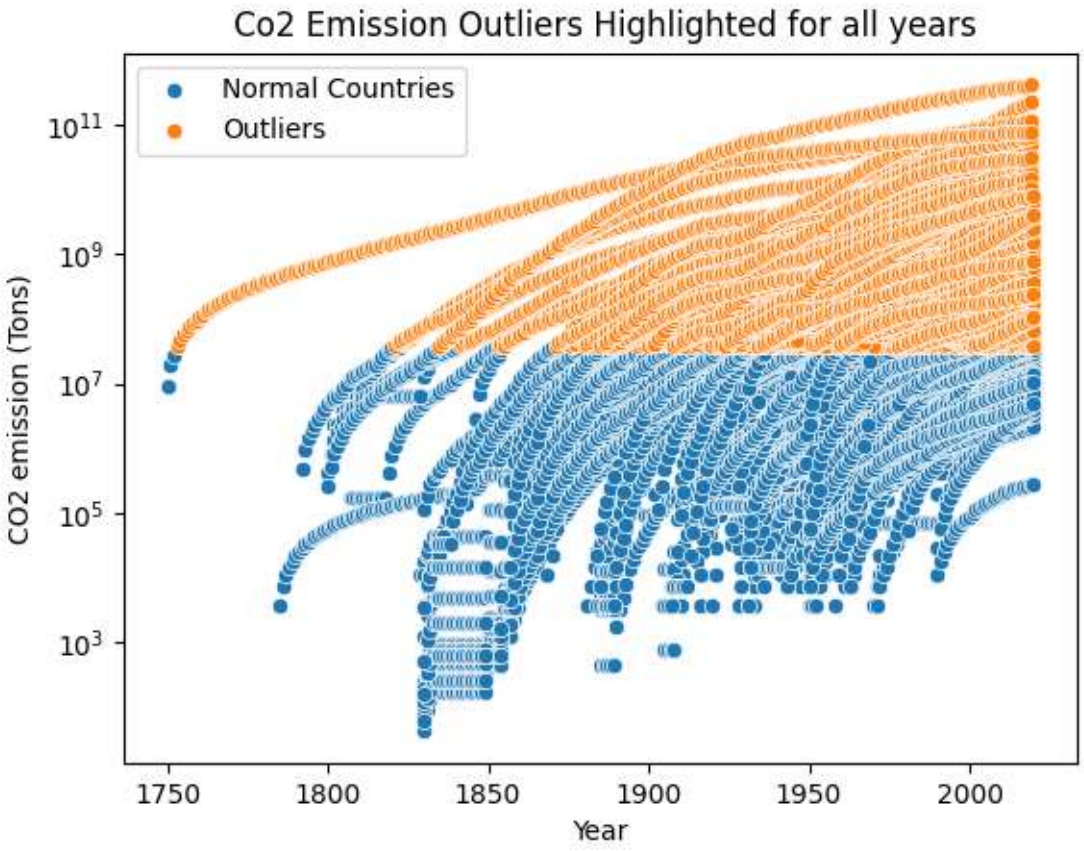

Figure 9 Scatter plot showing normal points and outliers\n of Co2 Emission for all year .

Similarly, throughout the years 1750 to 2020 we can observe many outliers. These outliers divert from the original trend, due to some countries emitting higher emissions compared to those of similar population or area range.

2.2. Model Building

For task 1: Neural Network

A multi-layer Perceptron (MLP) classifier was used to perform a binary classification of CO2 emission class. The neural network learns non-linear relationship between numerical country-level features such as population, area density, and other features.

The dataset was first separated into features and label , then they were split into test and train sets to evaluate how well the model generalizes to unseen data and prevent overfitting. While splitting, the data were shuffled and the test size was set to be 20 % of the total data. Data was then scaled using standardScaler as this data is uniformly distributed and contains outliers.

The neural network consist of input layers equal to numbers of features , two hidden layers with 100 and 50 neurons respectively, and an output layer with ne neuron for binary classification.

ReLU activation was sued in the hidden layers to introduce non-linearity and improve learning efficiency. The output layer uses a sigmoid activation to produce class probability. The model was trained using the adam optimization algorithm, to minimizing the log loss objective function. Trainning configuration includes max iteration up to 500, random state for reproducibility, and stratify since our label has imbalanced distribution.

Model was then fitted and predicted for both test sets and train sets. The final evaluation was done using accuracy, precision , recall, f1-score, and ROC-AUC. Whose values for both train set an detest were nearly equal to one another.

Task 2:

Model 1: Logistic regression

Initially all the features were used to predict the classification variable Emission_class. The dataset were first split in to test and training sets. Then using sklearn the data were fitted into a LogisticRegression with max iteration up to 1000. After training , class labels for the test data was predicted and also probability was calculated for ROC-AUC evaluation.

Model 2: Decision Tree

Previously split datasets were used to fit the training data in to a DecisionTreeClassifier. Then the class labels for the test data were predicted while also calculating the probability of the classs labels for ROC_-AUC evaluation.

2.3. Model Evaluation

For Neural Network:

For train data,

- Accuracy: 99.94%, overall prediction are almost all correct
- Precision: 78.9%, high emission prediction were usually right
- Recall: 61.2 %, missed some high emission cases
- F1-score: 0.69, moderate balance between precision and recall
- ROC-AUC: 0.9998, excellent ability to separate low vs high emission classes.

For test data,

- Accuracy: 99.94%, overall prediction is almost all correct, could be influence of imbalanced class
- Precision: 87.5%, high emission prediction were usually right, very few false positive
- Recall: 58.3 %, still missed some high emission cases
- F1-score: 0.70, slightly better balance between precision and recall
- ROC-AUC: 0.9999, near perfect ability to separate low vs high emission classes.

For Logistic Regression:

- Accuracy: 91.18%, overall prediction is nearly all correct,
- Precision: 1.2%, extremely low, almost all positive emission prediction were wrong
- Recall: 1.0, all high emission cases were correctly predicted
- F1-score: 0.024, since precision is very low balance between precision and recall ends up being low as well.
- ROC-AUC: 0.97, near perfect ability to separate low vs high emission classes.

For Decision Tree:

- Accuracy: 99.99%%, overall prediction is all correct,
- Precision: 1.0, extremely high, all positive emission prediction were right
- Recall: 91.66%, almost all high emission cases were correctly predicted
- F1-score: 95.65%, since both precision and recall are very high, balance between precision and recall is high as well.
- ROC-AUC: 95.83%, near perfect ability to separate low vs high emission classes.

2.4. Feature Selection:

Feature selection was done using the wrapper method and RFE with logistic regression and Decision Tree. Using the RFE most important features for the prediction of the Emission Class were filtered out. They were then visualized in a bar graph by the features and their ranking. The selected features were 'Year', 'CO2 emission (Tons)', 'CO2_per_capita', 'Population(2022)', 'Area', 'Density(km2)'. Final selected features:

Logistic regression:['Year', 'CO2 emission (Tons)', 'Population(2022)', 'Area', 'Density(km2)']

Decision Tree:['Year', 'CO2 emission (Tons)', 'Population(2022)', 'Area', 'Density(km2)']

2.5. Hyperparameter Optimization

By using GridSearchCV hyperparameter optimization was for both logistic and decision tree classification. K fold value of 5 was used and f1 was used for scoring , for both logistic and decision tree. Through this process hyperparamater values of 0.01 for logistic regression and max depth of 5 for decision tree were achieved.

3. Result and Conclusion

	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.999908	0.923077	1.0	0.96	1.00
Decision Tree	1.000000	1.000000	1.0	1.00	1.0

Table 1 metric evaluation before feature selection and hyperparameter optimization

	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.893542	0.010292	1.000000	0.020374	0.979852
Decision Tree	0.997878	0.333333	0.916667	0.488889	0.957317

Table 2 metric evalutaion after feature selection and hyperparameter optimization

It can be observed that the values after optimization feature selection are more realistic compare to the values before tunning and feature selection. Given our imbalance class data where the count of 1 values are less than 70 while 0 is in large more than 50000. This obtained evaluation is more accurate and realist in relation to used dataset.

Also, the derived emission class was derived from co2 per capital values by using a threshold of 3 which doesn't match the set threshold value by world bank which is 5 tons per person. Because of the high zeros in the co2 emission values the max co22 per person as only around 6.5. And since outliers such as values from pre and early industrial years differed heavily from that of after industrialization i.e after 1950 , to address this imbalance a threshold half of the max co2 per person was used.

For future works to minimize said imbalance we can remove the data of pre and early industrialized years ,and also research a more valid and logical threshold to differentiate the emission class for the train sets. Also more complex and better algorithms can be applied for better results.

4. Discussion

In regard to model performance, accuracy , precision, recall, F-1 score, and ROC-AUC were used to measure the model performance. Both logistic and decision tree performed rather too exceptionally despite high imbalance class in the dataset.

The features obtained from feature selection include "Year", "CO2 emission (Tons)", "CO2_per_capita", "Population(2022)", "Area", "% of World", "Density(km2)", while hyper parameters of 0.01 for logistic regression and max depth of 5 for decision tree were obtained. Using these parameters and selected features a farr more realistic evaluation of the models were obtain. While accuracy, recall, and ROc-AUC remained nearly same before and after tuning and feature selectin, a huge difference in precision and f1-score was observed that related more to the data and gave more realistic results than the exaugurated result before tuning and feature selection.

The low scores for precision and f1-score emerged due to the imbalance class in the dataset. Since there were only aound 65 cases of high emission class in the dataset of more than 50000 data. Such skewed dataset can lead to model to bias towards the majority class in the predictions but the dataset itself is just very imbalance. As such, for future works it is better to recommend that the dataset be minimized to include equal distribution of both classes such that the distribution will be less skewed. Another way to remove the limitation can be minimizing the year range from 1750 to start from 1950 which will automatically remove the 0 values from the early industrialized years.

5. References

1. Primary Dataset: CO2 Emission by countries Year wise (1750-2022)
2. Dataset for getting accurate population data: Global CO_2 & Greenhouse Gas Emissions (1750–2024)
3. Dataset used for getting missing co2 emission values: Annual CO2 Emissions (1750 - 2022)