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 predict country-level CO_2 emissions using regression techniques, based on the "CO_2 emission by countries year wise (1750-2022)" dataset. The data required preprocessing to handle missing and inaccurate values, and exploratory data analysis (EDA) was performed to understand trends and distributions. Three predictive models were implemented: Ridge Regression, Decision Tree Regression, and a Neural Network (MLP) regressor. Model performance was evaluated using MSE, RMSE, MAE, and R^2 metrics. Feature selection was conducted using the wrapper method with Recursive Feature Elimination (RFE), and hyperparameter tuning was performed via GridSearchCV to optimize model performance. The results indicate that the Decision Tree model outperforms Ridge Regression and Neural Network, capturing non-linear patterns and skewed distributions effectively. The study demonstrates the importance of careful preprocessing, feature selection, and hyperparameter tuning in improving predictive accuracy for CO_2 emissions.

Contents

Table of Figures:

1. Introduction

1.1. Problem Statement

The goal of this report is to predict “CO2 emission (Tons)” variable using regression techniques.

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.

1.1. Objective

The objective of this analysis is to build a predictive regression model that estimates the “CO2 emission (Tons)” 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.

2.2. 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 .

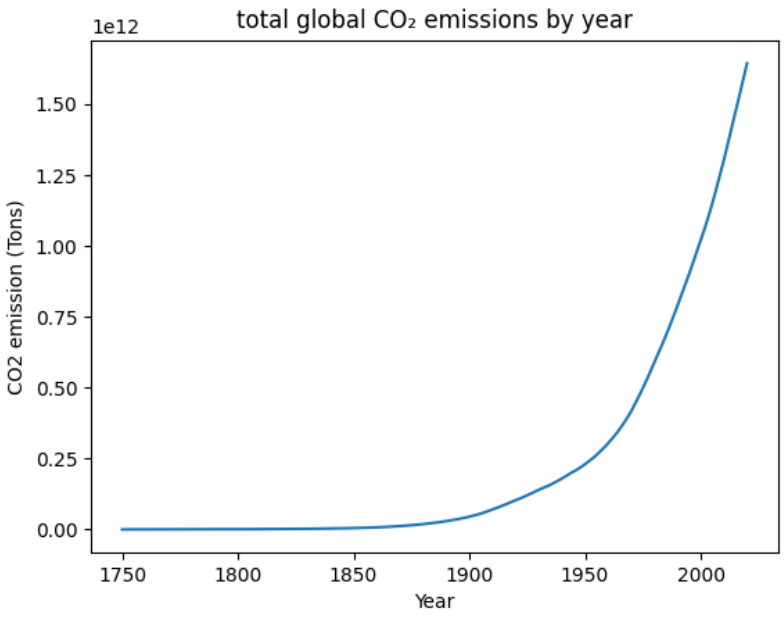

Figure 1Total co2 emission by years

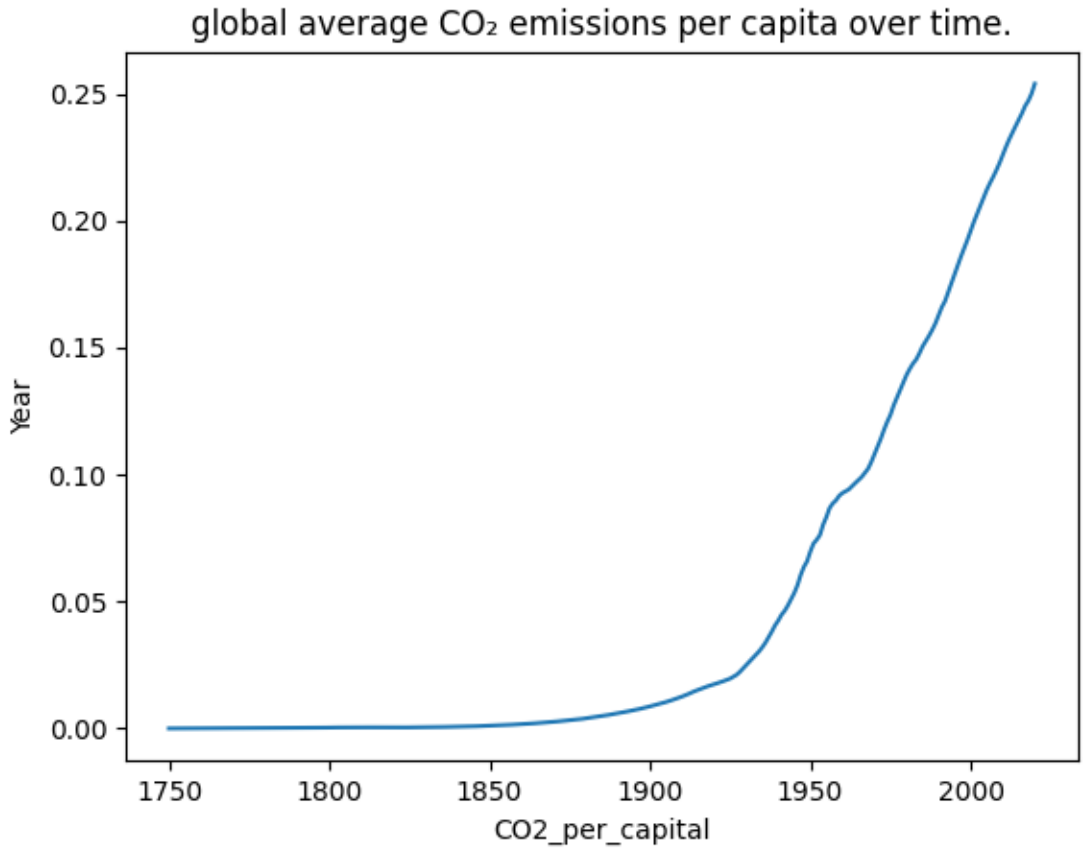

Figure 2 line chart showing the global average CO_2 emissions per capita over time.

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.

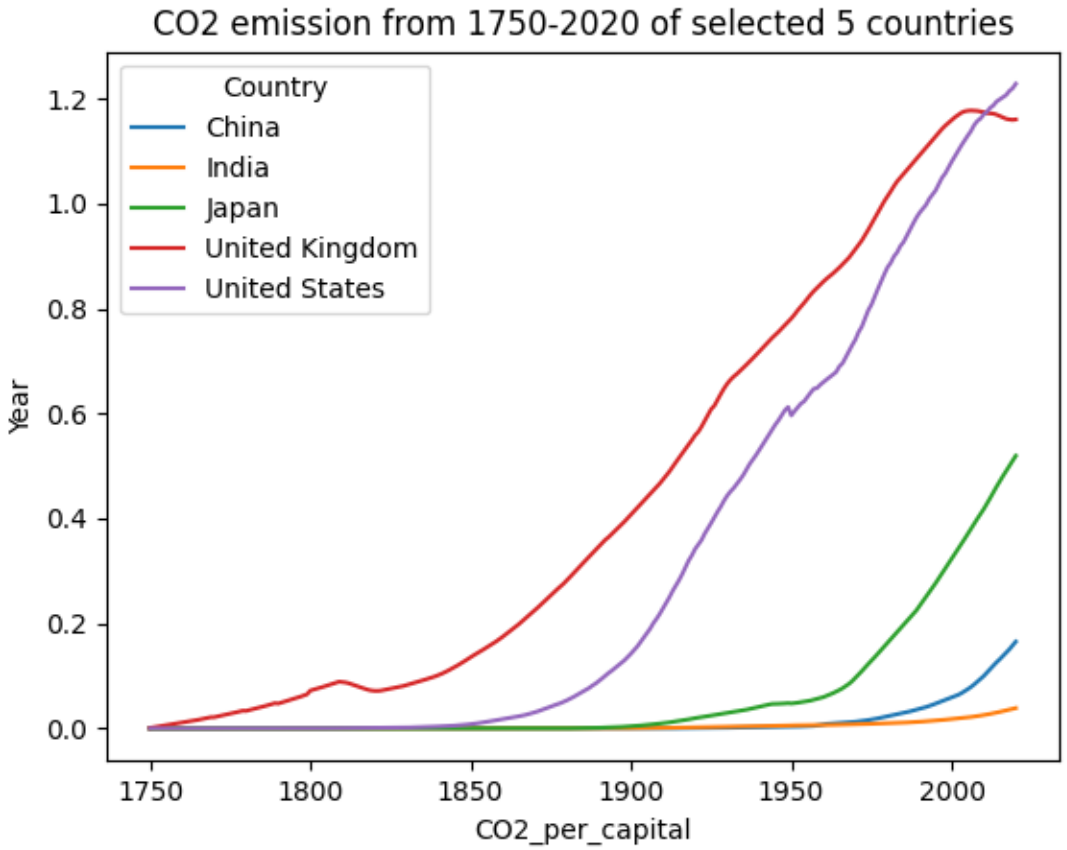

Figure 3line chart showing CO2 emission from 1750-2020 of selected 5 countries.

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.

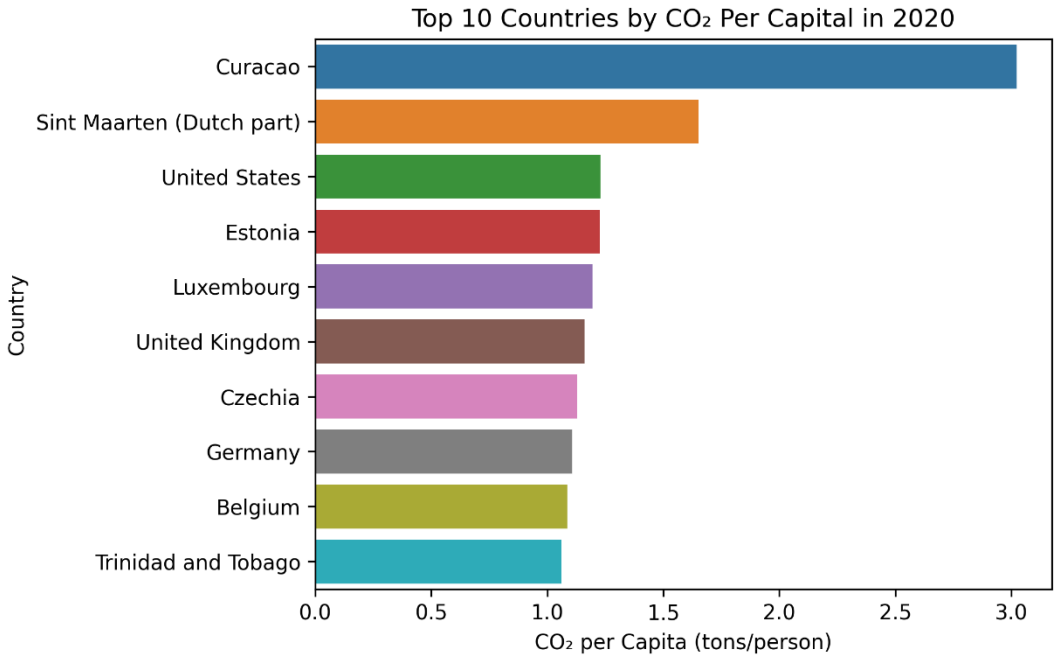

Figure:4 Bar chart showing top 10 Countries by CO2 per capital in 2020

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.

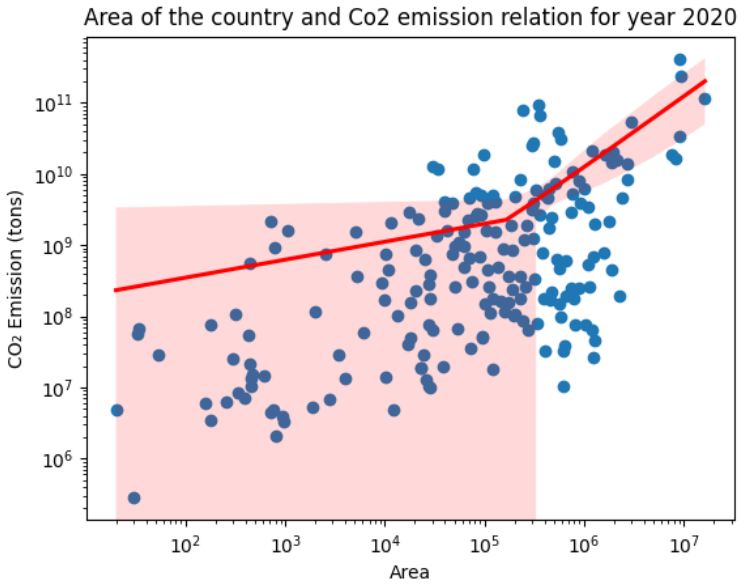

Figure 5 Scatter plot showing the relation between Area and the CO2 emission of a country for the year 2020

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.

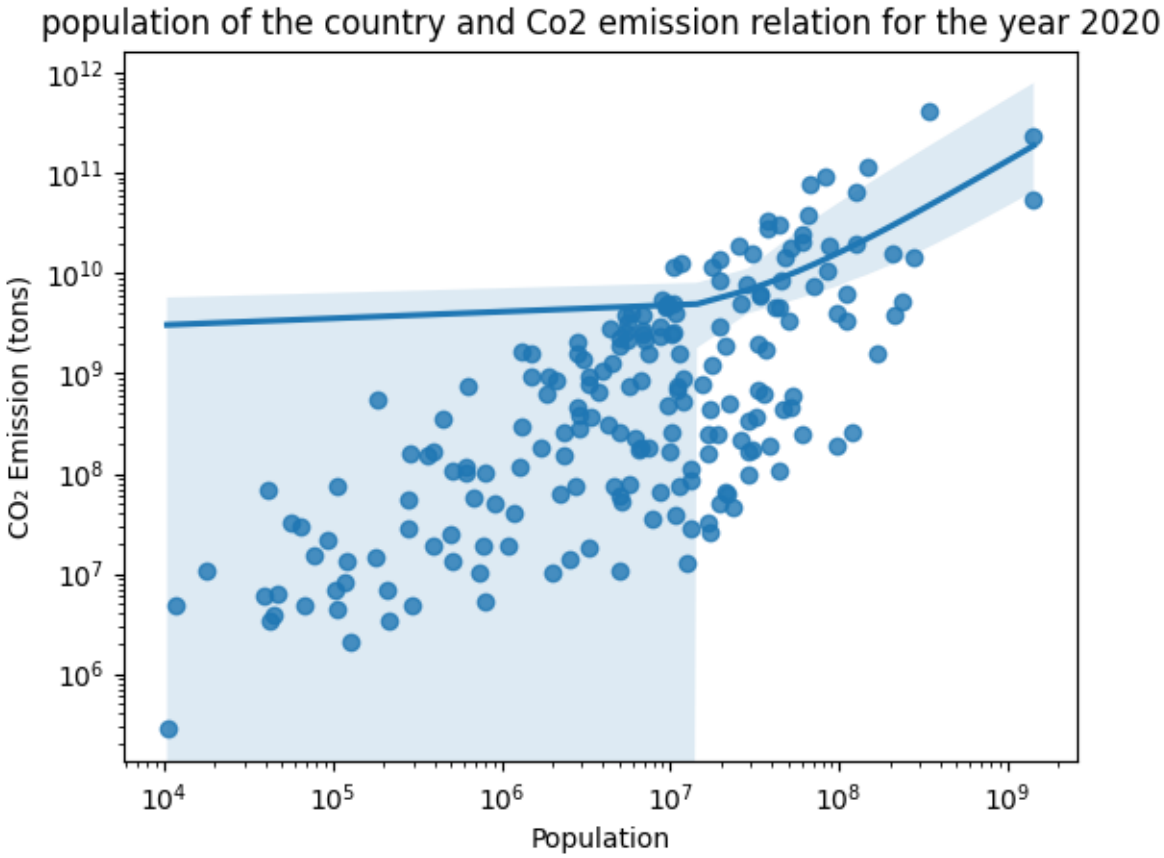

Figure 6 Scatter plot showing the relation between population and the CO2 emission of a country for the year 2020

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.

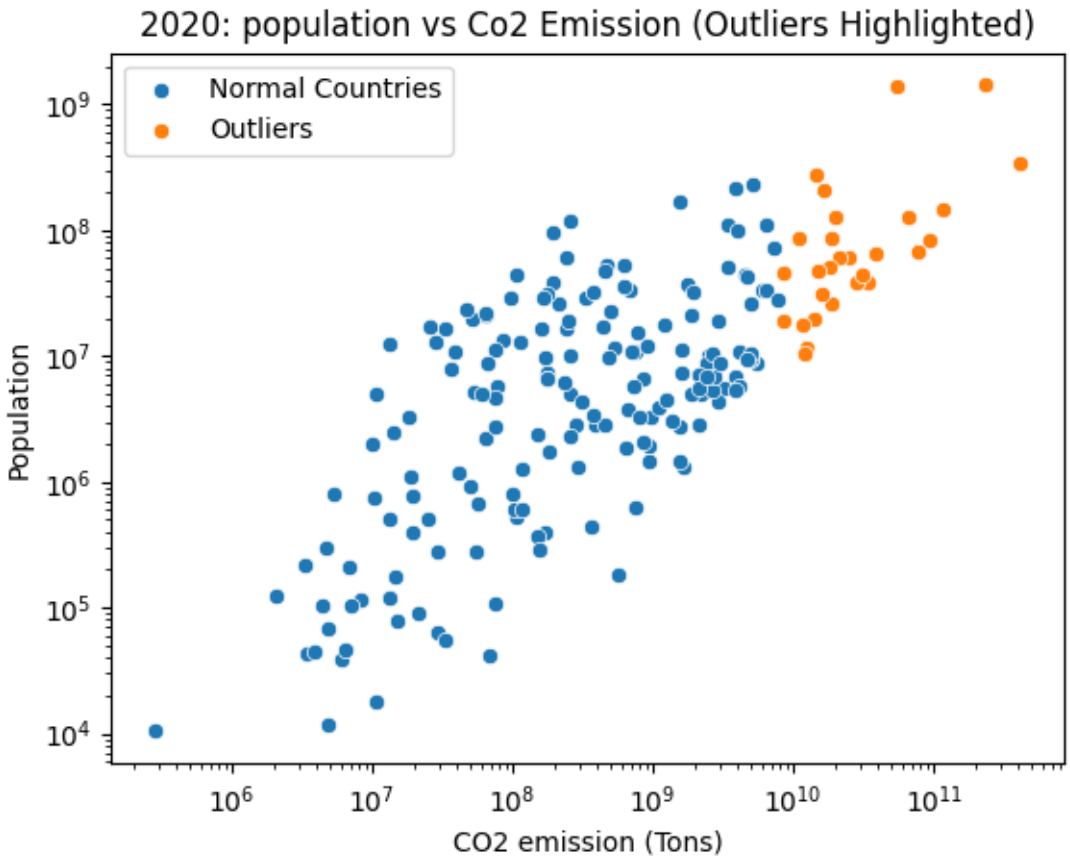

Figure 7 Scatter plot showing normal points and outliers\n among population vs Co2 Emission for year 2020.

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.

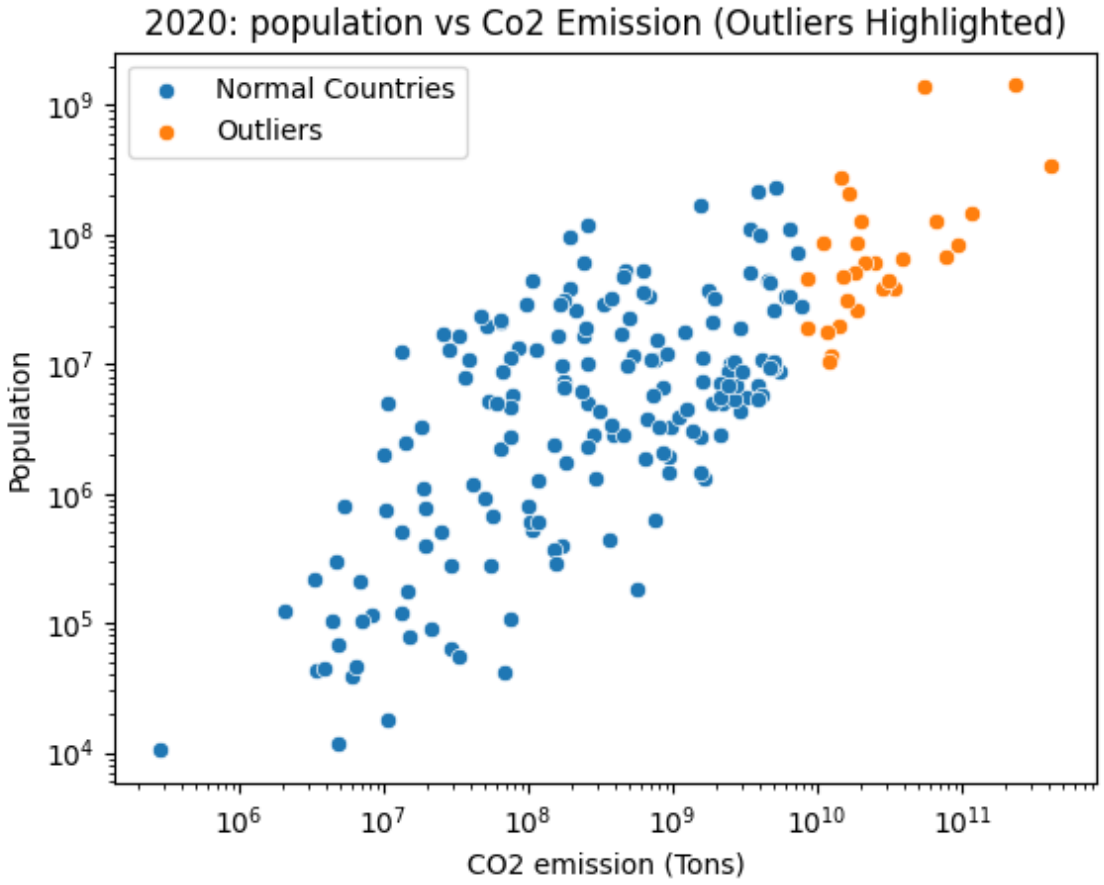

Figure 8 Scatter plot showing normal points and outliers\n among population vs Co2 Emission for year 2020.

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.

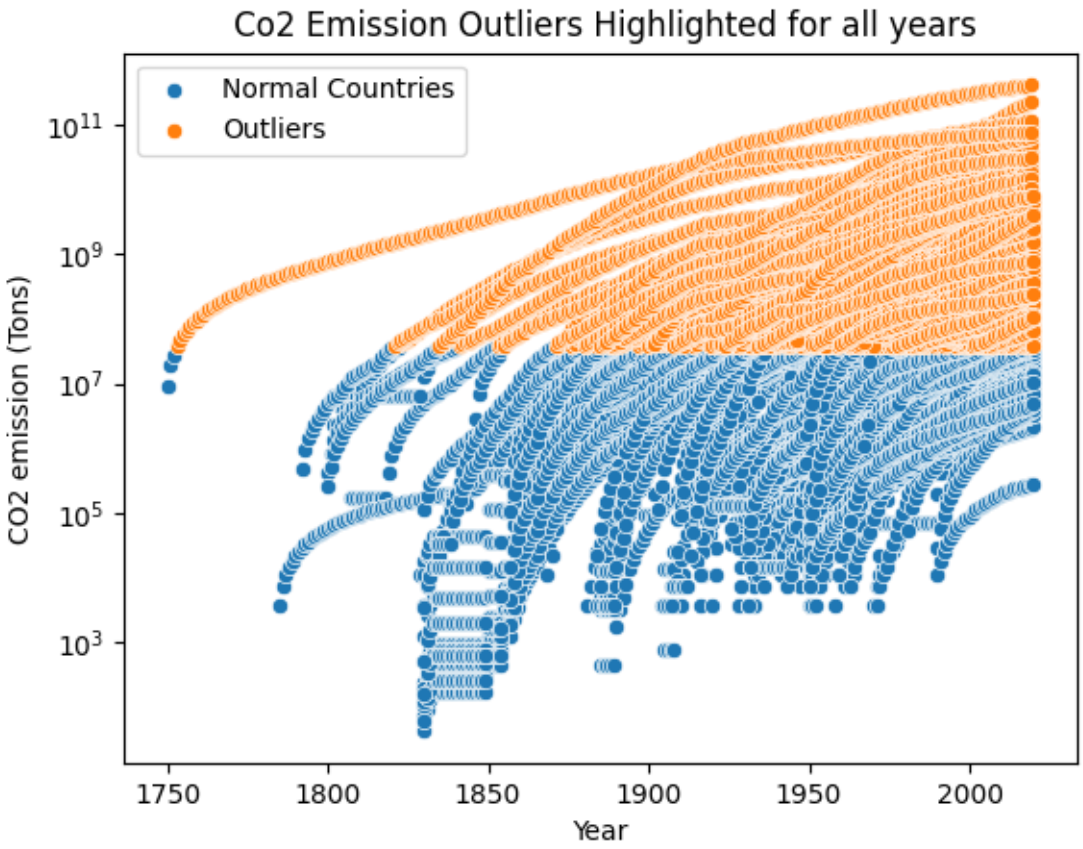

Figure 9 Scatter plot showing normal points and outliers\n of Co2 Emission for all year .

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.

2.3. Model Building

A multi-layer perceptron (MLP) regressor was used to predict continuous CO2 emission values. The neural network learns non-linear relationships between numerical country-level features such as population, area, density, and other relevant features.

The dataset was first separated into features (X) and target (y), then split into training and test sets to evaluate how well the model generalizes to unseen data and to prevent overfitting. While splitting, the data were shuffled, and the test size was set to 20% of the total data.

The data were then scaled using StandardScaler, as the features have different ranges and contain outliers, which helps the neural network converge faster.

The neural network consists of an input layer equal to the number of features, two hidden layers with 100 and 50 neurons respectively, an output layer with one neuron to predict the continuous CO2 emission.

ReLU activation was used in the hidden layers to introduce non-linearity and improve learning efficiency. The output layer uses a linear activation to produce continuous values.

The model was trained using the Adam optimization algorithm, minimizing the mean squared error (MSE) loss function. Training configuration included a maximum of 500 iterations and a random state for reproducibility.

The model was then fitted and predicted for both the training and test sets. The final evaluation was done using regression metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R^2 score, with results for both training and test sets being nearly equal, indicating good generalization.

Task 2:

Model 1: Ridge regression

Initially, all features were used to predict the continuous CO2 emission values. The dataset was first split into training and test sets. Then, using scikit-learn, the training data were fitted into a Ridge regression model with regularization parameter alpha (default or tuned) to prevent overfitting. After training, the model was used to predict CO2 emission values for the test set. A visual representation of the predicted values and the actual values proven to have no linear relation with one another. To captur this non linear relation decision tree regression was used as the next model.

Model 2: Decision Tree

The same training and test datasets were used to fit a Decision Tree Regressor. The model learned to map the features to the continuous CO2 emission values. After training, predictions were made for the test set and the model performance was evaluated using regression metrics (MSE, RMSE, MAE, R^2) to assess how well the model captures the variations in CO2 emissions across different observations.

2.4. Model Evaluation

For Neural Network:

For train data,

- MSE ≈ 9.49×10^{19} → the average squared error is extremely large, showing poor fit.
- RMSE ≈ 9.74 billion → typical prediction error is around 9.7 billion tons, very high.
- MAE ≈ 1.89 billion → on average, predictions are off by 1.89 billion tons.
- R^2 ≈ 0.103 (~10%) → the model explains only 10% of the variance, so it captures very little of the actual emission variation.

For test data,

- MSE :1.56×10^{20} , the squared errors between predicted and actual CO2 emissions are extremely large, indicating poor prediction.
- RMSE: 12.49 billion , the typical prediction error is around 12.5 billion tons, which is huge.
- MAE : 2.14 billion , on average, predictions are off by 2.14 billion tons, very high.
- R^2 : 0.0984 (~10%) , the model explains only 10% of the variance in CO2 emissions, meaning it performs very poorly.

Extremely large errors in prediction of co2 emissions.

For Ridge Regression:

- MSE : 1.36×10^{20} , on average, the squared errors between predicted and actual CO2 emissions are extremely large.
- RMSE: 1.17×10^{10} , typical prediction error is around 11.7 billion tons, which is huge.
- MAE: 2.23×10^{9} , on average, predictions are off by 2.2 billion tons.
- R^2 : 0.21, the model explains only 21% of the variance in CO2 emissions, so it's performing poorly.

Extremely errors can be seen and the predictions are off by a large values.

For Decision Tree:

- MSE : 1.02×10^{17}, on average, the squared errors between predicted and actual CO2 emissions were smaller than model 1.
- RMSE: 3.19×10^{8}, typical prediction error is around 319 million tons, which is far lower than model 1.
- MAE: 36.2 million , on average, predictions are off by 36 millions tons.
- R^2 : 0.9994, the model explains 99.94% of the variance in CO2 emissions, so it's performing is better thank model 1.

Comparatively, much better performance, with very small prediction errors is very highly accurate.

2.5. Feature Selection:

Feature selection was done using the wrapper method and RFE with Ridge regression model. Using the RFE most important features for the prediction of co2 emissions 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)', 'Population(2022)', 'Area', 'Density(km2)',"%of world".

Feature selection was done using the wrapper method and RFE with Decision tree regression model. Using the RFE most important features for the prediction of co2 emissions 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)', 'Population(2022)', 'Area', 'Density(km2)',"%of world".

In both selestions process, co2_per_capital and emission_class ranked higher but because they were directly or indirectly derived from the label variable , both of these variables were removed to prevent data leakage.

2.6. Hyperparameter Optimization

By using GridSearchCV hyperparameter optimization was for both ridge regression and decision tree regression. K fold value of 5 was used and neg_mean_squared_error was used for scoring , for both logistic and decision tree. Through this process hyper paramater values of 0.01 for ridge regression and max depth of None for decision tree were achieved.

3. Result and Conclusion

By using the hyper parameters and features obtained from feature selection and optimization, the evaluation of the 2 models are as follow

Metrics	Ridge Regression	Decision tree
MSE	1.5675668017571555e+20	1.0467727806839931e+18
RMSE	12520250803.227367	1023119142.9564756
MAE	2589031904.918612	116513375.20779674
R2	0.09336943714879997	0.9939458006238393

It can be observe that the performance worsen slightly after feature selection/ tuning. This is likely due to the highly skewed dataset and non-linear relationship. Neural Network model also performs similar to the ridge regressor and it can be that the neural network is underfitted due to small dataset for high emission cases

Although, the performance decreased slightly the results are extremely accurate. Since decision tree can capture the non-linear relationship better than ridge regression. Similarly this also outperforms NN dramatically due to its rorbust to skewed and non-linear relationships.

The CO2 per capita values in the dataset exhibit a highly skewed distribution, with many countries reporting zero or very low emissions and only a few reporting high values. The maximum CO2 per person in the dataset is around 6.5 tons, which is lower than typical thresholds such as the World Bank's 5 tons per person used for classification purposes. Additionally, extreme outliers exist due to pre- and early-industrial years, which differ significantly from post-1950 values after industrialization.

For future works to minimize said skewedness 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

Through EDA, various insights were carried out. It was concluded that population and Area were not significant enough to be considered as a sole factor to determine the co2 emission of a country.

In regard to model performance, the models were evaluated using regression metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R^2 score. Both Ridge Regression and Decision Tree Regression were applied to predict the continuous CO2 emissions.

The features used for modeling included "Year", "Population(2022)", "Area", "% of World", and "Density(km2)". Features like CO2 emission (Tons) and CO2 per capita were excluded to prevent data leakage, as they are directly related to the target variable.

For hyperparameter tuning, Ridge Regression used an alpha of 0.01, while the Decision Tree was trained without a maximum depth constraint to allow it to fully capture non-linear relationships in the data.

Using the selected features and tuned hyperparameters, a much more realistic evaluation of the models was obtained. The Decision Tree achieved excellent predictive performance, with very low errors and an R^2 close to 1, while Ridge Regression performed more moderately, capturing only part of the variance due to the non-linear nature and skewness of CO2 emissions in the dataset.

The skewed distribution of CO2 emissions, with a majority of low values and only a few extremely high values, influenced model performance. To improve regression results in future work, it is recommended to reduce the dataset to a more representative range, for example starting from 1950, which removes early-industrial zero values. It is also suggested to consider advanced models such as Gradient Boosting or Neural Networks that can handle skewed, non-linear data more effectively, and to explore further feature engineering to capture complex relationships without causing data leakage.

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)