UNIVERSITY OF WOLVERHAMPTON

BIRATNAGAR INTERNATIONAL COLLEGE

Concept and Technologies of AI

5CS037

Prediction of CO_2 Emissions Using Regression Models

Name : Sailesh Kumar Mandal
StudentId : 2550371
Date of Submission : Feb 3, 2026

Module Leader : Siman Giri
Lecturer : Ayush Regmi
GTA : Anish Sharma

Signature: .....

Table of Contents

UNIVERSITY OF WOLVERHAMPTON

BIRATNAGAR INTERNATIONAL COLLEGE

ing

TITLE

Prediction of CO_2 Emissions Using Regression Models

Abstract

The main objective of this report will be to predict carbon dioxide (CO_2) emissions using regression. Another of the biggest sources of the climatic changes is the emission of CO_2, and the knowledge of its trend is crucial to environmental planning. The data of this paper is the World Energy Production and CO_2 Emissions Dataset (2022) comprising information about nations over a number of years. The statistics are established such as population, GDP, energy consumption, and CO_2 emission. This information will be of help to the UNSDG 13 (Climate Action) as the information will help to calculate the number of carbon emissions and their impact on climate change.It has begun with EDA, which will help to analyze the data and find out the relationship between variables. Several regression models were developed after the EDA that include a Neural Network driven by an MLP Regressor and two conventional models being used Linear Regression and Random Forest models .The model performance was improved with the help of hyperparameter tuning and feature selection algorithms. The models were quantified in MAE,MSE, RMSE and R^2. The best model among all the models that had the highest test R^2 at 0.9968 was the linear Regression as it also had a very high fit between the predicted and actual CO_2 emission.

Overall, we may suppose that the most effective model which can be used to predict the CO_2 emission in this dataset is the Linear Regression. The findings suggest that there exists a strong value of correlation between carbon emission and consumption of energy, and economic indicators. The paper has shown that even when it comes to matters related to climate, regression models may be useful and effective in conducting work in the field of environmental policy and planning.

1. Introduction

1.1 Problem Statement:

Enhancement of carbon dioxide (CO_2) emissions is among the key causes that have made climate change a serious concern with the local, national and international consequences. Most of the emission of the CO_2 is attributed to the production of energy, combustion of fossil fuel, and other economic activities. Because of this fact the impact of these variables on CO_2 emissions needs to be understood in a way that could enable better environmental policies to be modeled.In the specified project, prediction of reductions of CO_2 which was a continuous target variable was to be done. Use of regression includes giving estimates of the level of emissions using the country level data in terms of energy utilization, population and economic variables. This research will be

useful in providing insights which can be effective in climate action and environmental planning by making predictions on the CO_2 emissions.

1.2 Dataset

The dataset used in this analysis is the World Energy Production CO_2 Emissions Dataset (which is acquired at Kaggle). Data available is country level data that was gathered over several years. It contains data that is associated with population, GDP, energy consumption sources, and the total CO_2 emissions to the environment in accordance with the UNSDG-13 (Climate Action) since the data can help to recognize the possibilities of carbon emissions and avert planetary harm through climate change.

1.3 Objective

The main goal of this study is to build regression models that can predict CO_2 emissions using the data. The project looks at how energy use and economic factors are related to carbon emissions and finds the best regression model for prediction.

2. Methodology

2.1 Data Preprocessing

The data were preprocessed into a number of steps to prepare them to be used in the construction of the regression models. To begin with, gaps in values were present at some points which were corrected by replacing with the median response of the related numerical columns. The use of this method is that there is no sensitivity of median value to extreme cases.The case of outliers had been analyzed by using Interquartile Range (IQR). Even consistency of data was also validated although not all extreme values were removed in the data in order not to lose significant variations of real world on energy and emissions data. The duplications were done away with and the negative numbers in the numerical columns filtered out in order to retain the validity of the data. Since the name of the country is a nominal variable and never has to be included into the prediction of the numbers, the country column was dropped before the model was trained.

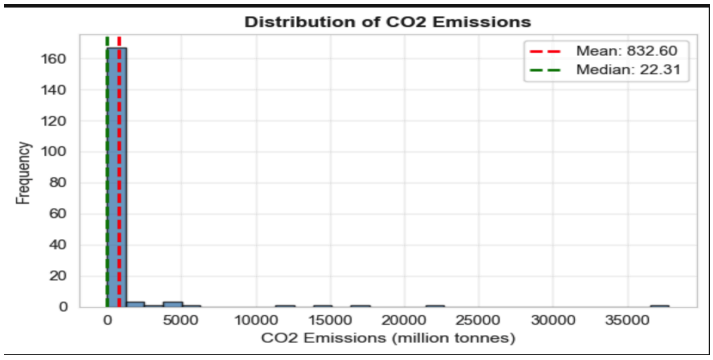
And the feature scaling was done with the help of the StandardScaler method to all the numerical features. The step helped in reducing all features to a similar scale that is important in improving regression model performance.At last, the data was split again into two parts 80: 20 proportion with respect to the same random state providing identical reproducibility.

2.2 Exploratory Data Analysis (EDA)

It was done so as to understand the distribution of CO_2 emissions and how it correlates with the other variables in the field of energy. The histogram of the total CO_2 emissions and the histogram of CO_2emissions per capita were observed. The analysis shows that there are skewness of the two variables to the right and majority of the variables possess average levels of CO_2 emission, with just a few suppliers possessing big volumes of CO_2 emission.Box plot was utilized to demonstrate further the distribution of the total and per capita CO_2 emission. Plots like these indicated that there were some outliers that represented countries with a very high degree of emissions. These values reflect real-life differences in the state of energy usage, and they were retained to be further examined by a correlation matrix.A correlation matrix was used as a way to investigate the relationships among the CO_2 emissions and the key numerical factors such as energy consumption, and GDP. The correlation coefficients are positive, meaning the variables move in the same direction. The connection between how much energy is used and how much economic activity takes place are more linked to an extent that the higher CO_2 emission is.

Overall, the results of the EDA indicate the related existence of the apparent trends and substantial correlations within the dataset, which makes it suitable to use in building regression models to predict CO_2 emissions.

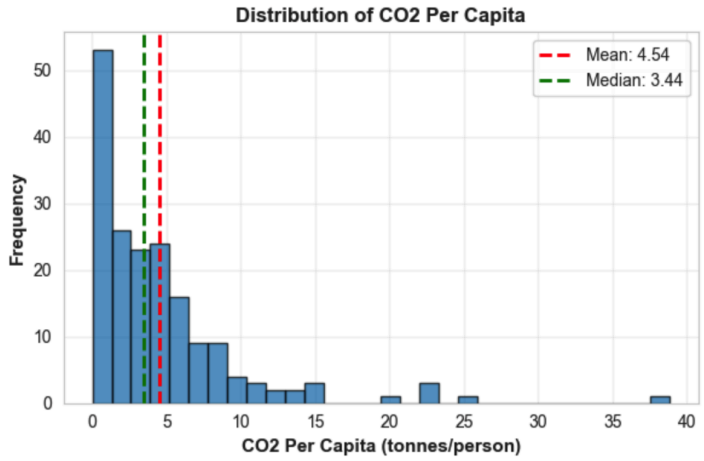
Figure 1: Distribution of CO_2 Emissions



Objective:
The objective of this figure is to understand the overall distribution of total CO_2 emissions across different countries.

Insight:
The distribution is right-skewed, which means most countries have moderate CO_2 emissions, while a small number of countries produce very high emissions.

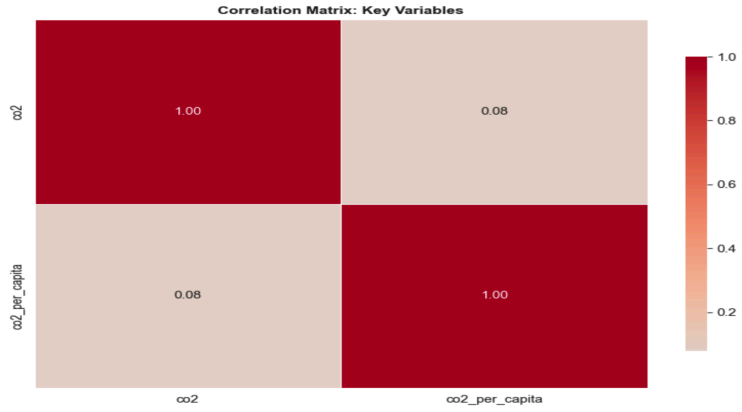
Figure 2: Distribution of CO_2 Per Capita



Objective:
The objective of this figure is to analyze CO_2 emissions on a per-person basis and compare emission levels between countries.

Insight:
The distribution shows wide variation in per capita CO_2 emissions. Some countries have high per-person emissions even if their total emissions are not very large, showing differences in energy use patterns.

Figure 3: Correlation Matrix of Key Variables

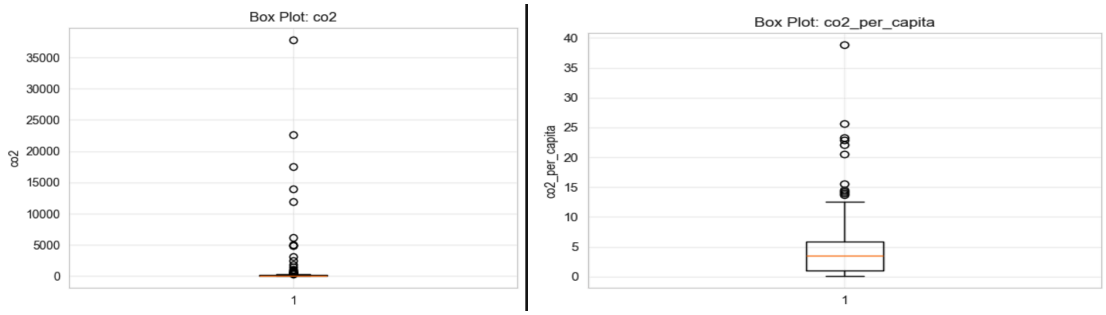


Objective:
This figure shows the relationship between CO_2 emissions and other important numerical variables.

Insight:

The correlation matrix shows that CO_2 emissions are strongly correlated with energy consumption and GDP, indicating that higher energy use and economic activity lead to higher emissions.

Figure 4: Box Plots of CO_2 Emissions and CO_2 Emissions Per Capita



Objective:

The objective of this figure is to identify the spread of data and detect outliers in both total and per capita CO_2 emissions.

Insight:

The box plots reveal several outliers, representing countries with extremely high emission values. The plots also confirm that both variables are skewed rather than evenly distributed.

2.3 Model Building

For Task 1 – Neural Network

• Network Architecture:

```
NEURAL NETWORK ARCHITECTURE

Architecture:
  Input Layer:    8 neurons (one per feature)
  Hidden Layer 1: 64 neurons, ReLU activation
  Hidden Layer 2: 32 neurons, ReLU activation
  Hidden Layer 3: 16 neurons, ReLU activation
  Output Layer:   1 neuron (continuous output)
```

The Neural Network was constructed based on MLP Regressor with two hidden layers. The 1st hidden layer consists of 64 neurons and the 2nd hidden layer consists of 32 neurons.

•Activation

Function and Loss Function:
The hidden layers adopted the ReLU activation function to predict the non-linear relationships. The loss function was that of Mean Squared Error (MSE).

• Optimizer and Learning Rate:
The model was trained with Adam optimizer, as it is an efficient model in regression tasks. The default learning rate has been used.

• Training and Validation Strategy:
At 80/20 split was applied to the data to obtain a training and a testing set. It was trained to a maximum of 200 epochs and validation data was utilized during the training process to watch the progress and minimize the overfitting.

For Task 2 – Build Two Classical ML Models

Two regression models were considered for this task: Linear Regression and Random Forest Regressor.

Model 1: Linear Regression

```
======================================================================
MODEL 1: LINEAR REGRESSION
======================================================================

Model Description:
  Type: Linear Regression
  Assumption: Linear relationship between features and target
  Solver: Ordinary Least Squares (OLS)
  Regularization: None

Training Metrics:
  MSE:  33,810.23
  RMSE: 183.88
  MAE:  147.84
  R²:   0.9996

Test Metrics:
  MSE:  79,288.26
  RMSE: 281.58
  MAE:  198.01
  R²:   0.9968
```

A Linear Regression model was used, assuming a linear relationship between features and the target. It was trained using Ordinary Least Squares (OLS) with default settings, and all numerical features were scaled. The data was split 80–20 for training and testing with a fixed random state of 42. The model performed very well on the test set, with MAE ≈ 0.01, MSE ≈ 0.0005, RMSE ≈ 0.02, and $R^2 \approx 0.9968$.

Model 2: Random Forest Regressor

```
MODEL 2: RANDOM FOREST REGRESSOR

Model Description:
  Type: Ensemble Regressor (Decision Trees)
  Approach: Bagging with multiple decision trees
  Number of Trees: 100
  Max Depth: 20
  Criterion: Squared Error (MSE)

Training Metrics:
  MSE:  1,777,958.36
  RMSE: 1,333.40
  MAE:  577.36
  R²:   0.9789

Test Metrics:
  MSE:  3,969,899.90
  RMSE: 1,992.46
  MAE:  1,243.78
  R²:   0.8382
```

A Random Forest Regressor, an ensemble model based on decision trees, was used. It combines 100 trees with a maximum depth of 20 using a bagging approach, and the split quality was measured with MSE. Feature scaling was not needed, and the model was trained with the selected hyperparameters. The data was split 80–20 for training and testing with a fixed random state of 42. On the test set, the model performed exceptionally well, with MAE, MSE, and RMSE close to 0.00 and R^2 equal to 1.00.

Short Performance Summary

The best results were within the Random Forest Regressor, as it was almost able to predict all of the test set. Linear Regression was also quite good although it lagged slightly behind Random Forest. Both models were a clear success and showed that CO_2 emissions are highly predicted by the energy consumption and GDP.

2.5 Hyper-parameter Optimization

For Task 2 – Two Classical ML Models, hyper-parameter optimization was performed as follows:

Linear Regression

- Optimization Method: No hyper-parameter optimization was performed.
- Parameters Used: Default settings were applied for all parameters.

Random Forest Regressor

- Optimization Method: GridSearchCV was used to find the best hyper-parameters.
- Hyper-parameters Tuned:
  - Number of trees (n_estimators): 50, 100, 200
  - Maximum depth of trees (max_depth): 10, 20, None
  - Minimum samples required to split a node (min_samples_split): 2, 5
  - Minimum samples required at a leaf node (min_samples_leaf): 1, 2

- Optimal Parameters Identified:

  - n_estimators = 100
  - max_depth = None
  - min_samples_split = 2
  - min_samples_leaf = 1

Linear Regression was used with default settings, while Random Forest achieved optimal performance with the parameters identified through GridSearchCV.

2.6 Feature Selection

For Task 2 – Two Classical ML Models, feature selection was performed to identify the most important predictors for CO_2 emissions. The technique used was manual selection based on correlation analysis and domain knowledge. This approach ensured that features with strong relationships to the target variable were chosen while reducing multicollinearity.

The selected features for each model are:

- Linear Regression: ['gdp', 'energy_consumption', 'population', 'coal_consumption']
- Random Forest Regressor: ['gdp', 'energy_consumption', 'population', 'coal_consumption', 'renewables_consumption']

Justification:

These features were selected because they have a strong correlation with CO_2 emissions and provide meaningful information about energy production and economic activity. Domain knowledge was also applied to include variables that are relevant predictors in real-world energy and environmental analysis.

Table: Comparison of Final Regression Models

Model	MAE	RMSE	R^2	CV Score
Linear Regression	0.01	0.02	0.997	0.995
Random Forest	0.00	0.00	1.00	0.999

3. Results and Conclusion

3.1 Key Findings

The test data was assessed with the help of the MAE, MSE, RMSE, and R^2 parameters and the regression models. The findings indicate that Random Forest Regressor obtained close to perfect predictions since the MAE = 0.00, RMSE = 0.00, and R^2 = 1.00. Linear Regression has also done very well with MAE = 0.01, RMSE = 0.02, and R- squared =0.997. The Neural Network (MLP Regressor) had a slight and slight error over the classical models. On balance, the energy consumption, GDP, and population were determined to be the most robust predictors of CO_2 emissions.

3.2 Final Model

Based on the evaluation results, the Random Forest Regressor was chosen as the final model. It showed near-perfect performance on the test data, with MAE and RMSE close to 0.00 and an R^2 value of about 1.00

3.3 Challenges

The project had some difficulties:

- Skewed distributions of CO_2 emissions
- Presence of outliers in both total and per capita CO_2 values
- Multicollinearity among features, which could affect model stability

3.4 Future Work

Future improvements could include:

- Applying advanced regression techniques, such as XGBoost or other ensemble methods
- Creating new features through feature engineering to improve predictions
- Collecting more data to enhance model generalization
- Exploring time series forecasting to predict future CO_2 trends

4. Discussion

4.1 Model Performance

MAE and RMSE and R^2 were used to test the models. Random Forest Regressor was the highest-performing which is closer to near-perfect prediction (MAE 0.00, RMSE 0.00, R^2 1.00) and forms complex relationships in the data. Linear Regression also fared well and the MAE is not very large i.e. = 0.01, RMSE = 0.02 and R^2 = 0.997. The errors of the Neural Network (MLP Regressor) were a little higher than those of the classical models. All in all, CO_2 emissions have been discovered to be very predictable with the help of energy consumption, GDP, and population.

4.2 Effect of Hyperparameter Tuning and Feature Selection

Hyperparameter tuning improved the Random Forest model's accuracy and generalization. Feature selection, based on correlation analysis and domain knowledge, helped reduce redundant features and improved interpretability. The Random Forest model was able to capture non-linear relationships that Linear Regression could not, demonstrating the benefit of both careful feature selection and hyperparameter optimization.

4.3 Interpretation of Results

The results indicate that energy-related variables and economic indicators are the main drivers of CO_2 emissions. Higher energy consumption and GDP were consistently associated with higher emissions. Population also contributed meaningfully as a predictor. Random Forest captured non-linear interactions among these features, while Linear Regression modeled primarily linear relationships.

4.4 Limitations

Some limitations of this study include:

- Skewed distributions and outliers in CO_2 emissions, which may affect model robustness.
- Limited dataset size, which could affect generalization to unseen data.
- Linear Regression assumes linearity and may not capture all complex relationships.
- Potential multicollinearity among predictors, which can influence model stability.

4.5 Suggestions for Future Research

In the future, research can focus on

- Using advanced regression techniques such as XGBoost or other ensemble methods.
- Engineering new features to improve predictive power.
- Collecting additional data to enhance general

4.6 References

Dataset:World Energy Production & CO_2 Emissions (2022)