AIDI-1002 - AI Algorithms

Group 7

Nirav Patel (100870539)

Anuj Patel (100904232)

Parth Patel (100901183)

Preet Patel (100897482)

Darshan Ghoghari (100897771)

Introduction

Greenhouse gas emissions are a significant contributor to global climate change, and understanding the patterns and trends in these emissions is crucial for developing effective policies and strategies to mitigate their impact. In this report, we will analyze a dataset of greenhouse gas emissions by country and year, and explore the trends in emissions and emissions per capita over time.

Data Description

The Global Carbon Project provided the dataset for this research, which comprises projected yearly greenhouse gas emissions for various nations and areas worldwide. The data span the years 1750 to 2021 and contains emissions data for individual countries, groupings of countries (e.g., the EU-28), and non-country organizations. (e.g. "International Aviation").

The dataset contains 63104 rows and 11 columns, with the following variables:

- Year: The year for which the emissions data is reported
- Country: The name of the country or region for which emissions data is reported
- ISO Code: The ISO code for the country or region
- Total Emissions: The estimated annual greenhouse gas emissions in millions of metric tons of CO2 equivalent
- Population: The population of the country or region in the given year
- Coal - how much coal is used from 1750 to 2021 in the given nations
- Gas - This shows how much Gas is used over the given time span
- Oil - The oil usage of the country in the given years
- Cement - The forecasted cement usage in the given regions
- Flaring - The Projected Flaring usage for given countries
- Other resources - Remaining sources used in the given countries

Data Preprocessing

The data was cleansed and preprocessed before any analysis to guarantee its quality and uniformity. The following procedures were carried out:

- The data was reduced to include just the years 1750-2021, and the field "Unnamed: 0" was removed.
- The data was organized by country and year, and total emissions were determined for each country-year combination.
- The emissions data were combined with a second dataset containing population statistics by nation and year.
- The statistics on emissions were normalized by dividing them by the population, yielding emissions per capita.

Exploratory Data Analysis

we did exploratory data analysis after preprocessing the data to get insights into the patterns and trends in greenhouse gas emissions throughout time. To assist us in exploring the data, we generated many visualizations, including:

- A line graph shows CO2 emissions over time and total greenhouse gas emissions.
- A bar chart depicts the distribution of all resource types and how much it was used.
- A bar chart states the given all nations and total co2 emissions.
- Another bar chart represents years and CO2 emissions.
- A bar chart illustrates the importance of all resources and the total without cement.

Training and testing data

On this dataset, we perform sklearn library train_test_split in order to split data.

- The dataset is initially preprocessed by scaling the numerical characteristics and removing the columns 'Country,' 'ISO 3166-1 alpha-3,' and 'Total' from the independent variable X. 'Total' is the value of the target variable y.

- The train_test_split function is then used to divide the preprocessed dataset into training and testing sets, with the training set containing 80% of the data and the testing set containing 20%. To guarantee that the split is repeatable, the random_state option is set to 42.

- The shape of the train and test set print using shape attributes, then splitting the train and test data for building and evaluating the machine learning model.

Training ML models

We use three ml algorithms such as a few algorithms such as Linear regression, Random forest regressor, and MLPregressor.

- Linear regression is a class in the sklearn.linear_model package. The fit approach is used to train the model on the training sets X_train and y_train.
- Using the prediction technique, the model is then utilized to generate predictions on the test set X_test. y_pred_lr stores the expected values.
- The sklearn.metrics module's mean_squared_error and r2_score functions are then used to assess the model's performance on the test set. The mean squared error (MSE) is the average squared difference between the target variable's expected and true values. The coefficient of determination (R2) quantifies the proportion of the variance in the target variable explained by the independent variables.
- The evaluation findings are then printed. The linear regression model's MSE and R2 values are presented. The MSE is a measure of the model's prediction quality, with a lower number indicating greater performance. The R2 value ranges from 0 to 1, with higher values suggesting a better fit. The greater the R2, the more variation in the target variable the model explains.
- Secondly, The RandomForestRegressor class is included in the sklearn.ensemble package. The fit approach is used to train the model on the training sets X_train and y_train. The random forest's decision trees are specified by the hyperparameter n_estimators, which is set to 100. To ensure consistency, the random_state parameter is set to 42.
- Using the predict technique, the model is then utilised to generate predictions on the test set X_test. The anticipated values are saved in the variable y_pred_rf.
- Last but not least, MLPRegressor is a class in the sklearn.neural_network package. The fit approach is used to train the model on the training sets X_train and y_train. The hidden_layer_sizes and max_iter hyperparameters are set at (50,50) and 1000, respectively. The hidden_layer_sizes parameter sets the number of neurons in each hidden layer of the neural network, while the max_iter parameter specifies the

maximum number of iterations required for the solver to converge. To ensure consistency, the random_state parameter is set to 42.

- The predict method is then applied to the model to generate predictions on the test set X_test. y_pred_mlp stores the expected values.
- The sklearn.metrics module's mean_squared_error and r2_score functions are then used to assess the model's performance on the test set. The mean squared error (MSE) is the average squared difference between the target variable's expected and true values.

Results

Several major facts concerning greenhouse gas emissions throughout time were uncovered by our analysis:

- From 1750 to 2021, global greenhouse gas emissions climbed continuously, with a few minor decreases in the mid-1990s and during the 2008 financial crisis.
- China, the United States, and India were the top three greenhouse gas emitters in 2017, with the European Union as a whole ranking fourth.
- When emissions are normalized by population, Qatar, Trinidad & Tobago, and Kuwait were the top polluters per capita in 2017, with the United States and Australia all placing well.
- Emissions per capita in many nations have typically decreased over time, while there is still substantial heterogeneity among countries and regions.

Conclusion

In conclusion, our analysis of greenhouse gas emissions data revealed some important Insights into the patterns and trends in emissions over time. Our findings highlight the need for continued efforts to check how much CO2 emission will be increased in the given regions. By understanding the factors driving emissions and identifying the usage of all resources for evaluating the increase of co2 emissions.