Predicting Co2 Emission by countries Using Machine Learning

Introduction

1.1 Project Overview

The project aims to predict CO2 emissions by countries using machine learning techniques. By analyzing historical data from 1960 to 2015, we aim to understand the relationship between CO2 emissions and various factors such as GDP ,fertility rates and lot more. This analysis will help in identifying trends and patterns that can be used for more accurate future predictions.

1.2 Objectives

i. Data Collection and Preprocessing: Gather and preprocess historical data on CO2 emissions, GDP, fertility rates, and other relevant factors for different countries from 1960 to 2015.
ii. Exploratory Data Analysis (EDA): Perform EDA to understand the data distribution, identify patterns, and detect any anomalies or outliers.
iii. Feature Selection: Identify the most significant factors influencing CO2 emissions to build an effective predictive model.
iv. Model Development: Develop various machine learning models to predict CO2 emissions and compare their performance.
v. Model Evaluation: Evaluate the models using appropriate metrics to determine their accuracy and robustness.
vi. Deploying: Deploying the model using Flask

Project Initialization and Planning Phase

2.1 Define Problem Statement

The primary goal of this project is to predict the CO2 emissions of countries from 1960 to 2015 by considering various socio-economic and environmental factors, such as GDP, fertility rate and etc.. Understanding these emissions patterns is crucial for developing effective policies to mitigate climate change.
CO2 prediction problem statment Report:Problem Statment

2.2 Project Proposal

This project aims to predict CO2 emissions by countries using machine learning techniques, analyzing historical data from 1960 to 2015 to understand the relationship between CO2 emissions and factors like GDP and fertility rates. By identifying trends and patterns, we will develop models for accurate future predictions, providing insights and recommendations for policymakers to effectively address climate change. The project will involve data collection, preprocessing, exploratory data analysis, model development, and future predictions.

Co2 prediction Project proposal Report:Project proposal

2.3. Initial Project Planning

The initial project planning phase involves defining the project scope and objectives, which include predicting CO2 emissions using machine learning techniques and identifying influencing factors such as GDP and fertility rates. Key steps include identifying and collecting historical data from reputable sources, particularly using datasets available on Kaggle. A detailed project timeline with specific milestones will be developed for each phase: data collection, preprocessing, exploratory data analysis, model development, and future predictions. Resource allocation, including personnel, software, and computational resources, will be determined to ensure the project's successful execution.

CO2 project planning Report:Project planning

Data Collection and Preprocessing Phase

3.1. Data Collection Plan and Raw Data Sources Identified

The data collection plan involves gathering a comprehensive dataset from Kaggle, which includes historical CO2 emissions, GDP, fertility rates, and other relevant factors for 247 countries from 1960 to 2015. The dataset has been verified to contain no null values, ensuring completeness. Additional relevant datasets from Kaggle will also be explored to enhance the analysis. All datasets will be downloaded, stored securely, and organized for easy access and processing.

Co2 Data collection Report:Data collection

3.2. Data Quality Report

The initial review of the Kaggle dataset confirms its high quality, with no missing or null values, which ensures data integrity. Consistency checks will be performed to verify that all data entries are accurate and uniform across different variables. Any discrepancies or anomalies found during this process will be addressed, and a detailed data quality report will be generated to document the findings and any corrective actions taken.

Co2 Data quality Report:Data Quality

3.3. Data Exploration and Preprocessing

Data exploration will involve statistical analysis and visualization to understand the distribution and relationships among variables. Preprocessing steps will include normalizing the data to standardize scales across different countries and variables, and encoding categorical data if necessary. Outliers will be identified and addressed appropriately to prevent skewing the analysis. This phase aims to prepare the dataset for effective machine learning model development.

Co2 Data Exploration and Preprocessing:Preprocessing

Model Development Phase

4.1. Feature Selection Report

The feature selection process involves identifying the most relevant variables for predicting CO2 emissions. Using techniques such as correlation analysis, mutual information, and feature importance from models like Random Forest, we will evaluate the contribution of each factor (GDP, fertility rates, etc.) to the target variable. The feature selection report will document the chosen features, the rationale behind their selection, and their expected impact on the model's performance.

Co2 Feature Selection Report:Feature Selection Report

4.2 Model Selection Report

Various machine learning algorithms will be considered for predicting CO2 emissions, including Linear Regression, Random Forest, and XG Boost etc... Each model's suitability will be evaluated based on criteria such as interpretability, complexity, and computational efficiency. The model selection report will summarize the evaluation process, comparing the strengths and weaknesses of each model, and justify the selection of the most appropriate model for initial training.

Co2 Model Selection:Model Selection

4.3. Initial Model Training Code, Model Validation, and Evaluation Report

Initial training of the selected model will be conducted using the preprocessed dataset. The training process will include splitting the data into training and testing sets and implementing the chosen machine learning algorithm. Model validation will be performed using techniques such as cross-validation, and evaluation metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared will be reported. The code used for training and the evaluation results will be documented in the model validation and evaluation report.

Co2 model Evaluation : Evaluation Report

Model Optimization and Tuning Phase

5.1. Hyperparameter Tuning Documentation

Hyperparameter tuning aims to optimize the model's performance by finding the best combination of hyperparameters. Techniques such as grid search and random search will be employed to explore the hyperparameter space. The tuning process and the rationale behind choosing specific hyperparameter values will be documented, detailing the steps taken to achieve the optimal model configuration.

5.2. Performance Metrics Comparison Report

After tuning, the model's performance will be compared across different hyperparameter settings and model configurations. Key metrics such as MAE, MSE,Adjusted R squared and R-squared will be used to assess improvements in model accuracy and reliability. The performance metrics comparison report will present a comprehensive analysis of the results, highlighting the configuration that yielded the best performance.

5.3. Final Model Selection Justification

Based on the results from the performance metrics comparison, the final model will be selected. The justification for this selection will include an analysis of the model's accuracy, generalizability, and computational efficiency. The final model selection justification report will provide a detailed explanation of why the chosen model is the best fit for predicting CO2 emissions, considering both performance and practical deployment considerations.

Co2 Model Selction after Tuning:[Model Selection]

Results

6.1. Output Screenshots

```
print(f'Random Forest Regression R2 Score: {score1}')
```

```
Random Forest Regression R2 Score: 0.9830788955337789
```

Advantages & Disadvantages

Advantages

i. Accurate Predictions: Utilizing machine learning techniques allows for precise and reliable CO2 emission forecasts.
ii. Data-Driven Insights: Analysis of historical data reveals critical trends and patterns, aiding policymakers in decision-making.
iii. Scalability: The model can be adapted and extended to incorporate additional countries and factors over time.
iv. Automation: Reduces the manual effort required for data analysis and predictions, increasing efficiency.

Disadvantages

i. Data Quality Dependence: The accuracy of predictions heavily relies on the quality and completeness of the input data.
ii. Complexity: Developing and tuning machine learning models can be complex and require specialized knowledge.
iii. Resource Intensive: Computational resources and time required for model training and tuning can be significant.
iv. Limited Interpretability: Some advanced machine learning models, such as neural networks, can be challenging to interpret and understand.

Conclusion

This project successfully developed a machine learning model to predict CO2 emissions by analyzing historical data from 1960 to 2015. By identifying key factors influencing emissions and utilizing advanced modeling techniques, we provided accurate predictions and valuable insights for policymakers. The results highlight the importance of data-driven approaches in addressing global environmental challenges and offer a foundation for future enhancements and applications.

Future Scope

Future work could involve expanding the dataset to include more recent data and

additional factors such as technological advancements and policy changes. Integrating more sophisticated models and ensemble methods could further improve prediction accuracy. Additionally, developing an interactive dashboard for real-time predictions and scenario analysis could provide more accessible and actionable insights for decision-makers.

Appendix

10.1. Source Code

This section will provide the complete source code used for data collection, preprocessing, model development, training, validation, and evaluation. The code will be well-documented and organized to facilitate understanding and reproducibility.

Co2 SourceCode :Source Code

10.2. GitHub & Project Demo Link

A link to the GitHub repository containing the project's source code, documentation, and any additional resources will be provided. Additionally, a link to a live project demo, if available, will be included to showcase the model's functionality and results.

Video DemoLink: Video Link

Github Repository:Github Link