Final Project Report

1. Introduction

1.1. Project overview

Our project predicts CO2 emissions by countries using machine learning. We analyze data on country names, indicator names (e.g., CO2 emissions in metric tons), and years to forecast emission trends. This model aids policymakers and researchers in understanding and addressing climate change more effectively.

1.2. Objectives

- Develop a predictive model for CO2 emissions.
- Analyze CO2 emission trends by country and year.
- Support climate policy and mitigation strategies.
- Improve understanding of emission-influencing factors.

2. Project Initialization and Planning Phase

2.1. Define Problem Statement

Accurately predicting CO2 emissions is difficult due to complex influencing factors. Our project aims to improve prediction precision by using machine learning to analyze country data and indicator values, providing more reliable forecasts for effective climate action.

2.2. Project Proposal (Proposed Solution)

We propose using machine learning to predict CO2 emissions by analyzing data on country names, indicator names (e.g., CO2 emissions in metric tons), and years. This approach will enhance prediction accuracy and provide valuable insights for climate policy and mitigation strategies.

2.3. Initial Project Planning

- Define project scope and objectives.
- Collect data on countries, indicators, and CO2 emissions.
- Preprocess and clean the data.
- Select machine learning algorithms.
- Train and validate the model.
- Evaluate model performance.
- Deploy the model for practical use.
- Monitor and maintain the model.

3. Data Collection and Preprocessing Phase

3.1. Data Collection Plan and Raw Data Sources Identified

- **Dataset Link**: World Development Indicators
- **Data Collection**: Access and download the dataset from Kaggle.
- **Data Components**:
  - **Country Names**: Identification of countries in the dataset.
  - **Indicator Names**: Categories of CO2 emissions and other relevant indicators.
  - **Year**: Temporal aspect for historical data on emissions.
- **Initial Data Review**: Inspect the dataset for completeness and relevance to project objectives.
- **Data Integration**: Combine data from different sources if needed, ensuring alignment with project requirements.
- **Data Storage**: Organize and store data for easy access and processing in subsequent stages.

3.2. Data Quality Report

The Dataset which we had a lot of attributes which weren't required for the project as it was a world bank dataset which showed the world development indicators .Since our project only required the ones related Co2 emissions it was necessary to filter out some of the other irrelevant indicators

3.3. Data Exploration and Preprocessing

- **Label Encoding**:
  - Encoded country names and indicator names to numerical values for model compatibility.
- **Univariate Analysis**:
  - Analyzed individual features to understand their distribution and key statistics.
- **Bivariate Analysis**:
  - Examined relationships between pairs of variables to identify correlations and dependencies.
- **Multivariate Analysis**:
  - Assessed interactions among multiple variables to uncover complex patterns and relationships.
- **Data Cleaning**:
  - Handled missing values, outliers, and inconsistencies to ensure data quality.
- **Feature Engineering**:
  - Created new features or transformed existing ones to enhance model performance.

4. Model Development Phase

4.1. Feature Selection Report

- **Selected Features**:
    - **Country Name**: Identifies the country for analysis.
    - **Indicator Name**: Specifies the category of CO2 emissions.
    - **Year**: Provides the temporal context for the data.
    - **Value**: Represents the CO2 emission metric.
- **Excluded Features**:
    - **Country Code**: Redundant with Country Name.
    - **Indicator Code**: Redundant with Indicator Name.

4.2. Model Selection Report

- **Models Tested**:
    - **Random Forest Regressor**: Provided the highest accuracy and best results.
    - **Ridge Regression**: Evaluated but with lower accuracy compared to Random Forest.
    - **Polynomial Regression**: Assessed but did not perform as well as Random Forest.
- **Chosen Model**:
    - **Random Forest Regressor**: Selected for its superior performance and accuracy in predicting CO2 emissions.

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

- **Random Forest Regressor**:
    - **R^2 Score**: 0.9896
    - **Notes**: Highest accuracy, indicating excellent performance in predicting CO2 emissions.
- **Ridge Regression**:
    - **R^2 Score**: 0.02462
    - **Notes**: Significantly lower accuracy, suggesting poor model performance.
- **Polynomial Regression**:
    - **R^2 Score**: 0.12333
    - **Notes**: Moderate accuracy but still inferior to Random Forest.
- **Conclusion**: Random Forest Regressor was selected for its superior performance in terms of accuracy and prediction capability.

5. ModelOptimization and Tuning Phase

5.1. Hyperparameter Tuning Documentation

We have tuned n_estimators, random_state, and n_jobs hyperparameters in the

Random Forest model. The optimal values for these hyperparameters are 10, 52, and -1, respectively

5.2. Performance Metrics Comparison Report

We have compared the Baseline Metric and Optimized Metric for the Random Forest model. The baseline metric was 0.8936, and after optimization, the metric improved to 0.9986

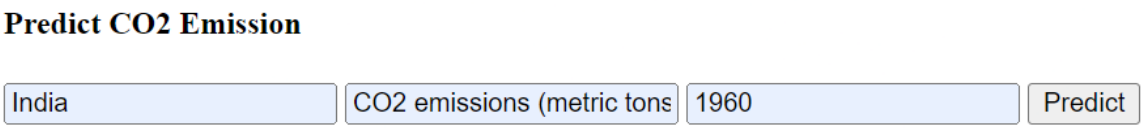
5.3 Final Model Selection Justification

- **Model Tuned**: Random Forest Regressor
- **Hyperparameters and Optimal Values**:
  - **n_estimators**: 10
  - **random_state**: 52
  - **n_jobs**: -1

Without modifying n_estimators, the model was overfit and was unable to handle values outside the dataset, resulting in a lower accuracy, therefore we picked Random forest (which has the highest accuracy in all the models we tested) and tuned it to have only 10 n_estimators.

6. Results

6.1. Output Screenshots



7. Advantages & Disadvantages

Advantages:

1. High Accuracy: The chosen Random Forest Regressor model achieved a very

high R^2 score of 0.9896, indicating excellent predictive performance.
2. Data-Driven Insights: The model provides valuable insights into CO_2 emission trends, helping policymakers and researchers make informed decisions.
3. Versatility: The approach can be applied to different countries and time periods, offering a flexible tool for emission analysis.
4. Feature Importance: Random Forest models can provide insights into which features are most important for predicting emissions, potentially highlighting key factors influencing CO_2 output.
5. Scalability: The model can be easily updated with new data, allowing for continuous improvement and long-term relevance.

Disadvantages:

1. Data Limitations: The model's accuracy is dependent on the quality and completeness of the input data. Missing or inaccurate data could affect predictions.
2. Complexity: Random Forest models can be more complex and less interpretable than simpler models, which may make it challenging to explain predictions to non-technical stakeholders.
3. Computational Resources: Training and running Random Forest models, especially with large datasets, can be computationally intensive.
4. Limited Causal Inference: While the model can identify correlations, it may not provide clear insights into causal relationships between factors and CO_2 emissions.
5. Potential for Overfitting: Without proper cross-validation and regularization, there's a risk of overfitting to the training data, which could reduce generalizability to new, unseen data.

8. Conclusion

This project successfully developed a highly accurate machine learning model for predicting CO_2 emissions by countries, with the Random Forest Regressor achieving an R^2 score of 0.9896. The model provides valuable insights for policymakers and researchers, enabling more informed climate change mitigation strategies. While highly effective, its performance depends on data quality and computational resources. This work demonstrates the potential of machine learning in addressing environmental challenges, offering a data-driven approach to shape climate policies. Future improvements could include expanding features and developing user-friendly interfaces, further enhancing its utility in combating climate change.

9. Future Scope

- Add more features (e.g., renewable energy, transportation data).
- Explore advanced techniques (e.g., deep learning).
- Include sub-national and more detailed temporal data.
- Develop real-time prediction capabilities.
- Implement robust cross-validation methods.
- Create user-friendly tools and dashboards.

10. Appendix

10.1. Source Code

https://github.com/Raghul727/Predicting-CO2-Emission-by-Countries

10.2. GitHub & Project Demo Link

Raghul.G: https://github.com/Raghul727/Predicting-CO2-Emission-by-Countries

Naveen: https://github.com/naveen-81229/Predicting-CO2-Emissions-by-Countries

Sudarshan: https://github.com/SudarshanPTS/Predicting-Co2-Emissions-of-Countries-Using-Machine-Learning-

Manivel: https://github.com/vel2004/Predicting-CO2-Emission-by-countries

Project Demo Link: https://drive.google.com/file/d/1-2xyWhAYEhIwVVzyF4fx9Bb0Q_R06pDd/view