



Project Proposal

Title

AI-Based Estimation of Carbon Emissions Using Machine Learning: A
Case Study on Saudi Arabia

Prepared By :

Name: Mohammad Wajeeh Bader

ID No: 251000462

Submitted in fulfillment of the requirements for the degree Master
of Science in Artificial Intelligence

At

College of Informatics

Midocean University

UAE

Supervisor Name: Dr. Amjad Qataish

College of Informatics, Midocean University, Comoros

1. Project Overview

This project aims to solve the problem of accurately predicting Saudi Arabia's annual CO₂ emissions using machine learning techniques. Currently, environmental monitoring relies heavily on manual calculations and traditional methods that lack precision and timeliness. This is important because Saudi Arabia is working toward Vision 2030 environmental goals and needs reliable tools to track emission progress and inform policy decisions.

The project will develop a machine learning framework that uses readily available national economic and energy indicators to predict CO₂ emissions. This approach will make emission forecasting more accessible, accurate, and cost-effective compared to complex satellite-based methods. The resulting tool will help environmental agencies, policymakers, and researchers monitor emission trends and evaluate the effectiveness of environmental policies.

By creating a practical, reproducible framework specifically tailored to Saudi data characteristics, this project will provide decision-makers with an evidence-based tool that supports environmental monitoring and contributes to achieving the country's sustainability objectives.

2. Project Goal

Primary Goal: Develop a model to predict carbon emissions in Saudi Arabia.

Specific Objectives:

- Create a machine learning model that achieves $R^2 \geq 0.80$ accuracy on test data
- Compare multiple algorithms to identify the most effective approach
- Establish guidelines for model selection in emission forecasting
- Deliver a reproducible framework for environmental monitoring applications

3. Approach

Data Collection

The project will use publicly available annual data from two primary sources:

- CO₂ Emissions Data: Our World in Data comprehensive database covering Saudi Arabia's historical emissions
- Economic and Energy Indicators: World Bank Open Data including up to 6 key features:
 - GDP per capita
 - Population and urbanization rate
 - Energy consumption per capita
 - Fossil fuel electricity percentage

- Renewable energy share
- Industrial production index

The data will cover the period from 1990-2024, subject to availability across all sources.

Data Processing

- Data Integration: Align temporal coverage and units across different data sources
- Quality Control: Handle missing values using documented interpolation methods
- Feature Engineering: Create lag features and 3-year moving averages to capture temporal patterns
- Data Validation: Implement quality checks and outlier detection to ensure data reliability

Model/Method Selection

The project will implement and compare multiple machine learning approaches:

- Baseline Models: Ridge and Lasso regression to establish performance benchmarks
- Primary Model: CatBoost regressor, selected for its effectiveness with small tabular datasets
- Optional Extension: XGBoost or LightGBM if time permits

Testing and Evaluation

- Validation Strategy: Time-based cross-validation using expanding window approach (5 folds)
- Hold-out Testing: Reserve final 5-7 years of data for ultimate performance assessment
- Performance Metrics: R^2 , MAE, RMSE, and MAPE to comprehensively evaluate model accuracy
- Analysis: Feature importance analysis, error pattern investigation, and model comparison

4. Project Steps and Timeline

Main Steps (12 Weeks Total)

Data Collection and Preparation (Weeks 1-3)

- Gather CO₂ emission data from Our World in Data
- Collect economic and energy indicators from World Bank
- Clean and integrate datasets from different sources
- Create temporal features and handle missing values

Model Building and Training (Weeks 4-7)

- Implement baseline models (Ridge and Lasso regression)
- Develop primary CatBoost model
- Perform hyperparameter optimization
- Train models using prepared dataset

Testing and Evaluation (Weeks 8-10)

- Execute time-based cross-validation
- Test models on hold-out data
- Compare performance across different algorithms
- Analyze feature importance and error patterns

Final Adjustments and Reporting (Weeks 11-12)

- Make final model improvements based on testing results
- Write comprehensive project report
- Organize code repository and documentation
- Prepare final presentation materials

5. Expected Results

This project aims to achieve the following outcomes:

- Improved accuracy in predicting carbon emissions for Saudi Arabia using machine learning techniques, achieving $R^2 \geq 0.80$ compared to traditional forecasting methods.
- New insights into which algorithms work best for emission prediction using national indicators, providing evidence-based guidance for similar environmental forecasting projects.
- A working tool that can be used for environmental monitoring, specifically supporting Saudi Arabia's Vision 2030 objectives and helping policymakers track emission reduction progress.
- Reproducible methodology and code that other researchers can apply to different regions or adapt for other environmental indicators.
- Clear performance benchmarks showing the effectiveness of different machine learning approaches for carbon emission forecasting in the Middle East region.

6. Resources Needed

Software and Tools

- Programming Environment: Python 3.9+ with data science libraries (pandas, numpy, scikit-learn, catboost)
- Computing Resources: Google Colab (free tier) or standard laptop
- Development Tools: Jupyter notebooks, GitHub for version control

Datasets

- Primary Sources: Our World in Data and World Bank Open Data (both free and publicly available)

- Backup Sources: EDGAR database for validation if needed

Other Support Required

- Expert Guidance: Regular check-ins with thesis advisor
- Computing Resources: Standard university computing facilities if needed

7. References

1. Cao, E. L. (2023). "Machine learning approaches for carbon emission prediction: A systematic review." *Frontiers in Environmental Science*, 11, 1285471.
2. Nassef, A. M., et al. (2023). "Enhanced Machine Learning Models for Predicting Carbon Dioxide Emissions in Saudi Arabia." *Sustainability*, 15(9), 7648.
3. Rouet-Leduc, B., & Hulbert, C. (2024). "Machine learning for carbon emissions prediction and mitigation strategies." *Nature Communications*, 15, 3801.
4. Chen, T., Guestrin, C. (2016). "XGBoost: A scalable tree boosting system." *Proceedings of the 22nd ACM SIGKDD International Conference*, 785-794.
5. Our World in Data (2024). "CO2 and Greenhouse Gas Emissions Database." Available: <https://ourworldindata.org/co2-and-other-greenhouse-gas-emissions>
6. World Bank (2024). "World Development Indicators: Energy and Environment." Available: <https://datacatalog.worldbank.org/search/dataset/0037712>
7. Saudi Green Initiative (2021). "National Environmental Strategy and Vision 2030 Goals." Available: <https://www.saudigreeninitiative.org>
8. Prokhorenkova, L., et al. (2018). "CatBoost: unbiased boosting with categorical features." *Advances in Neural Information Processing Systems*, 31.