# CO₂ Factor Forecasting – Report

## 1. Approach

The goal of this project was to forecast the CO₂ emission factor (gCO₂/kWh) for the next 7 days (hourly resolution), using historical renewable energy production and weather data as features.  
  
Steps Taken:  
1. Data Preparation:  
- Combined 4 years of hourly data (2021–2024) including renewable generation (solar, wind, biomass, electricity volume) and weather variables (radiation, temperature, wind speed, sunshine, precipitation).  
- Engineered time-based features (hour\_sin, hour\_cos, month, day\_of\_week) to capture daily and seasonal patterns.  
- Removed redundant or highly correlated variables (capacity columns).  
  
2. Feature Selection:  
- Applied stepwise regression and correlation analysis to retain the 12 most relevant predictors

3. Modeling Approach:  
- Trained multiple ML models: OLS, Ridge, Lasso, Random Forest, XGBoost (untuned and tuned), LightGBM, KNN, and MLP Neural Network.  
- Used 80/20 train-test split and 5-fold cross-validation for evaluation.  
- Hyperparameter tuning for XGBoost was performed using GridSearchCV.

## 2. Assumptions

- Weather and renewable generation strongly influence CO₂ factor, and past patterns are reliable predictors of future trends.  
- Data from trusted sources (electricity and weather datasets) are accurate and free of outlier-related errors.  
- Stationarity is not strictly required since tree-based models (XGBoost, RF) can handle trends and seasonality effectively.

## 3. Results

The following models were compared using metrics such as RMSE, MAE, and R² for both train and test sets. Cross-validation (CV RMSE) was also computed to ensure generalization performance.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | RMSE (train) | MAE (train) | R² (train) | MAE (test) | R² (test) | CV RMSE (mean) |
| OLS (Linear Regression) | 0.0269 | 0.0210 | 0.9209 | 0.0256 | 0.9108 | 0.0305 |
| Ridge | 0.0269 | 0.0210 | 0.9209 | 0.0256 | 0.9108 | 0.0305 |
| Lasso | 0.0271 | 0.0212 | 0.9194 | 0.0263 | 0.9032 | 0.0310 |
| Random Forest | 0.0050 | 0.0036 | 0.9973 | 0.0149 | 0.9705 | 0.0229 |
| XGBoost (untuned) | **0.0124** | **0.0092** | **0.9830** | **0.0127** | **0.9784** | **0.0214** |
| XGBoost (tuned) | **0.0133** | **0.0099** | **0.9806** | **0.0118** | **0.9816** | **0.0208** |
| LightGBM | 0.0106 | 0.0079 | 0.9877 | 0.0125 | 0.9792 | 0.0212 |
| KNN | 0.0155 | 0.0115 | 0.9735 | 0.0142 | 0.9725 | 0.0224 |
| MLP Neural Network | 0.0208 | 0.0161 | 0.9524 | 0.0179 | 0.9524 | 0.0239 |

The best performing model is XGBoost(tuned), since it has the highest R-square, RMSE, and MAE among the different models.

RMSE (train) and RMSE (test) are slightly different, showing that there are no overfitting issues.

Figure 1 Figure 2   
A graph showing a number of data

AI-generated content may be incorrect.A graph with blue and white bars

AI-generated content may be incorrect.  
A graph of a distribution of residuals

AI-generated content may be incorrect.Figure 3

Figure 2 shows that the most important features are onshore\_wind\_volume, solar\_volume, biomass\_volume.

The actual and predicted CO₂ factor values overlap closely, indicating that the tuned XGBoost model captures the temporal patterns effectively with minimal error. (Figure 1)

The residuals are centered around zero and approximately normally distributed, showing that the model's predictions are unbiased and consistent. (Figure 3)