University of Hertfordshire UH

Applying Linear Regression and Random Forests to Predict Air Quality: A Data Mining Study

Rahul Shah
23105760

Subtopic 1: Linear Regression

Subtopic 2: Decision Trees and Random Forests

Link to Google Colab notebook
Link to GitHub repository

Introduction

Air quality is a critical global concern, notably in my hometown of Kathmandu, Nepal, which recently topped the Air Quality Index (AQI) as the world's most polluted city (e.g., AQI 348 on April 3, 2025, per recent reports). This study applies two data mining techniques to predict carbon monoxide (CO) concentration using the UCI Air Quality Data Set.
Linear Regression models linear relationships between features and CO levels. Decision Trees split data based on features, while Random Forests ensemble multiple trees for robustness. The goal is to compare these subtopics.

Data Preprocessing

The Air Quality data set comprises 9357 hourly records with 13 numeric features (e.g., temperature, humidity, NO_2 levels) and CO concentration (CO(GT)) as the target. Preprocessing steps included:

- Dropping non-numeric Date and Time columns
- Converting missing values (marked as -200) to NaN, then imputing with column means using SimpleImputer. This addressed significant missing data, ensuring model compatibility despite real-world inconsistencies.
- Standardizing features with StandardScaler to ensure uniform scales.
- Splitting data into 80% training and 20% testing sets.

Table 1 summarizes the dataset:

Table 1

Attribute	Value
Samples	9357
Features	13
Target	CO(GT) (mg/m^3)
Missing Values	1683 (CO(GT))

Methodology

Linear Regression: Fits a linear equation to predict CO(GT) using all features, with no hyperparameter tuning.
Decision Trees & Random Forests: Decision Trees partition data by feature thresholds (random_state =42), while Random Forests average 100 trees (n_estimators=100) to enhance accuracy. MSE and R^2 evaluate performance.

Results and Comparison

Results are presented in Table 2 and visualized in Figures 1 and 2:

Table 2

Model	MSE	R^2
Linear Regression	0.333	0.811
Decision Tree	0.568	0.679
Random Forest	0.285	0.839

In figure 1, we can see the Predicted VS Actual CO(GT) for Linear Regression, Decision Tree and Random Forest.
Linear Regression's predictions (blue) exhibit moderate spread around the y=x line, with noticeable deviations at low and high CO(GT) values, reflecting its linear limitation.
Decision Tree (orange) displays the most scatter, with many points far from the y=x line, especially at mid-to-high CO(GT) values, indicating severe overfitting.
Random Forest (green) clusters are most tightly along y=x showing the fewest deviations and superior accuracy.

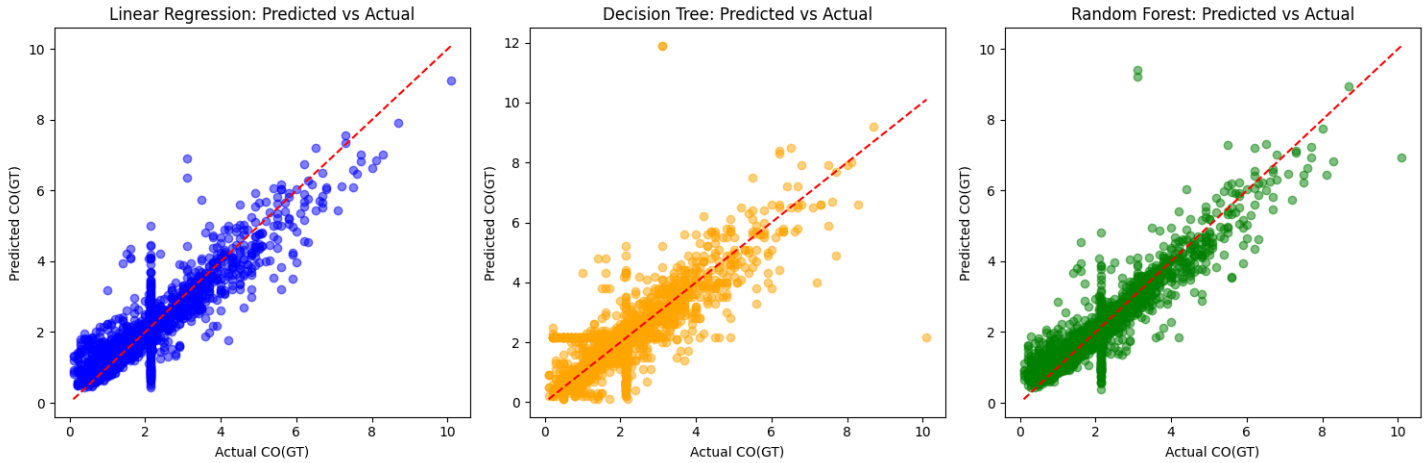

Figure 1

Figure 2 confirms Random Forest's lower MSE (0.28) and higher R^2(0.839) compared to Linear Regression (0.333, 0.811) and Decision Tree (0.568, 0.679). Within subtopic 2, Random Forest outperforms Decision Tree by 50% lower MSE and 24% higher R^2, as its ensemble of 100 trees reduces overfitting and variance, better capturing complex patterns like temperature-humidity interactions. Linear Regression's linear assumption limits its ability to model such interactions while Decision Tree's overfitting significantly hinders its performance.

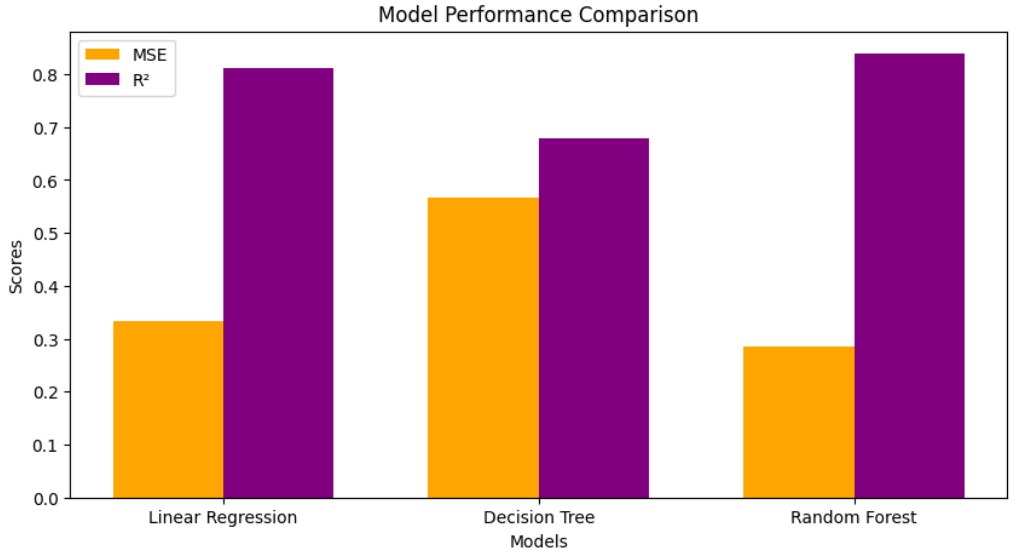

Figure 2

Conclusion

This study applied Linear Regression and explored Decision Trees and Random Forests to predict CO concentration in the UCL air quality dataset.

Within Subtopic 2, Random Forest outperforms Decision Tree by reducing overfitting through ensembling, achieving a 50% lower MSE (0.285 vs. 0.568) and 24% higher R^2 (0.839 vs. 0.679). Random Forest also surpasses Linear Regression (MSE: 0.333, R^2: 0.811), while Decision Tree underperforms both. Random Forest's accuracy suggests its potential for real-time forecasting in polluted regions, aiding policy decisions. Future work could explore feature engineering (e.g. interaction terms), time-series analysis for temporal trends, or integrating external data like traffic and industrial emissions. These findings underscore data mining's vital role in tackling global air quality issues through predictive modeling.

References

- UCL Machine Learning Repository. Air Quality Data Set. https://archive.ics.uci.edu/dataset/360/air+quality