Jakarta AQI Forecasting (2016–2021)

Final Report

Umar Hanif Abdul Aziz
Sabancı Üniversitesi

Spring 2025

1 Introduction

This project aims to forecast air quality in Jakarta using machine learning models based on historical AQI and weather data from 2016 to 2021. The objective was to implement both regression and classification models to predict next-day AQI values and categories, respectively.

2 Dataset Description

Air Quality Index (AQI) Data

Sourced from Kaggle, containing daily air pollution metrics and station readings.

Weather Data

Sourced from Visual Crossing API, including daily temperature, humidity, wind speed, UV index, cloud cover, etc.

Time Range

- January 1, 2016 to December 31, 2021
- Datasets were merged by the `date` column

3 Feature Engineering

Multiple features were engineered to enrich the dataset and improve model performance:

- **Lag Features**: `aqi_lag1`, `aqi_lag3`, `tempmax_lag1`
- **Rolling Averages**: `rolling_aqi_3`, `rolling_tempmax_3`

- **Date Features**: `month`, `day_of_week`, `is_weekend`
- **Interaction Features**: `tempmax × humidity`, `windspeed × cloudcover`
- **Categorical Transformation**: AQI and temperature binned into labeled categories

4 Machine Learning Models

Regression Task

- **Goal**: Predict next day's AQI value
- **Model**: Random Forest Regressor
- **Evaluation Metrics**:
  - MAE 22.5
  - RMSE 30.3

Classification Task

- **Goal**: Predict next day's AQI category
- **Model**: Random Forest Classifier
- **Evaluation Metrics**:
  - Accuracy 69%
  - Best performance on `Moderate` class (F1 0.77)

5 Results and Visualizations

Actual vs Predicted AQI

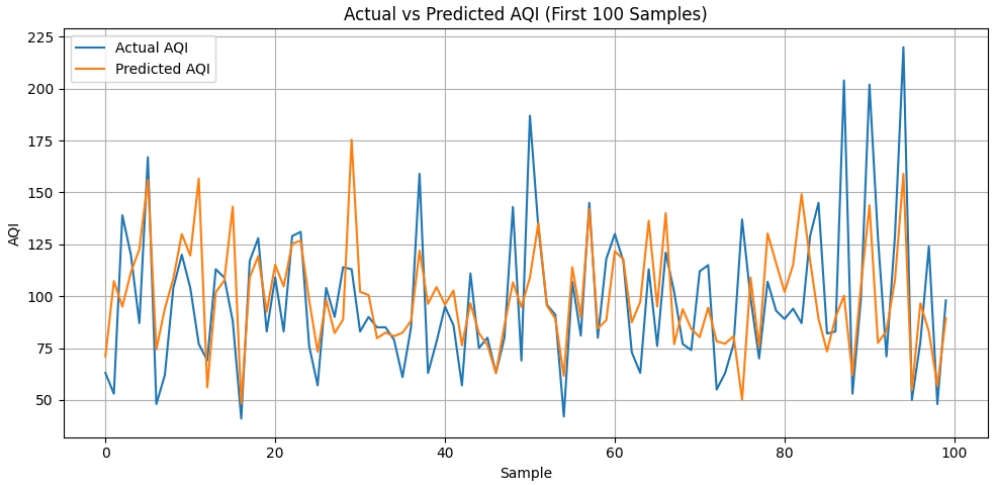

Figure 1: Comparison between actual and predicted AQI values (first 100 samples)

Feature Importance (Regression)

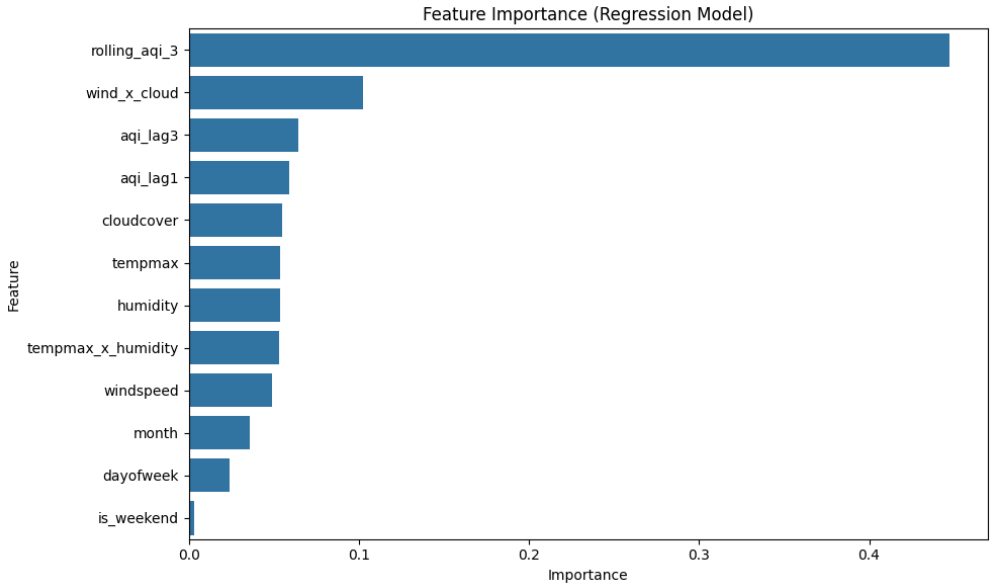

Figure 2: Top features used in regression model

Feature Importance (Classification)

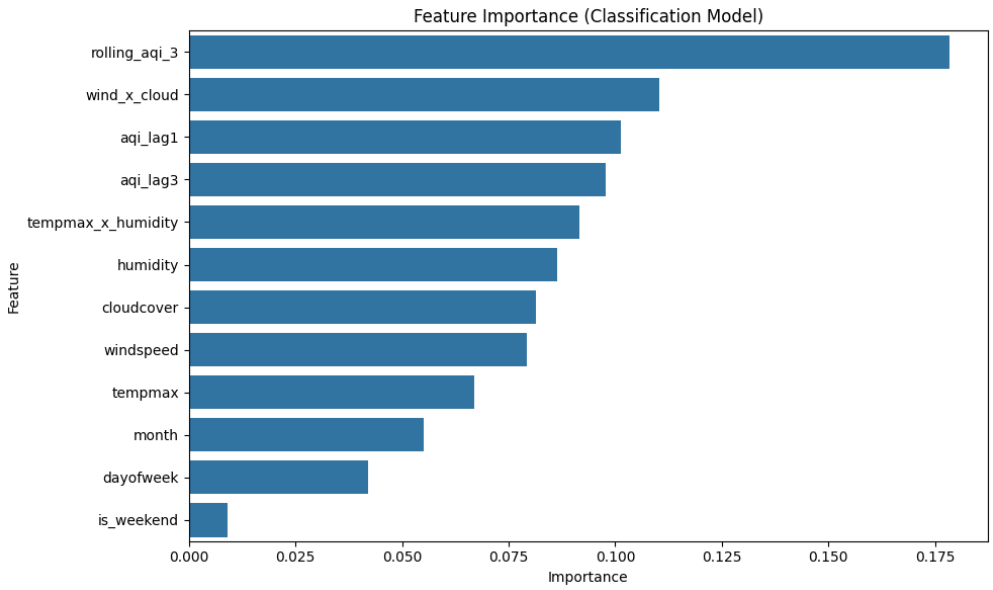

Figure 3: Top features used in classification model

6 Conclusion

The models demonstrated reliable performance in forecasting Jakarta's AQI. Historical pollution trends, especially rolling and lagged values, were the strongest predictors. Classification worked well for common AQI categories but struggled with underrepresented classes.

7 Limitations and Future Work

- Data imbalance affected classification for rare AQI levels
- Additional factors (e.g., traffic, public holidays, haze) could further refine the predictions

8 GitHub Repository

Project code and notebooks are available at:
`https://github.com/bonip155/jakarta-aqi-forecasting`