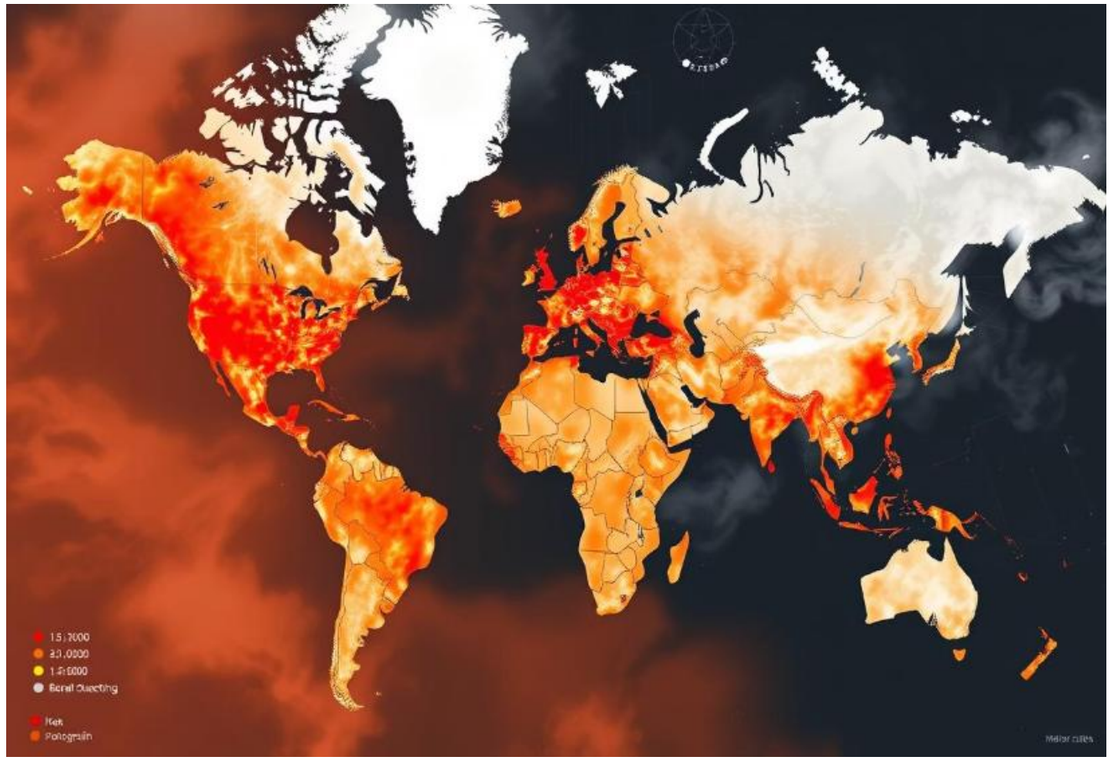
Air Quality Index (AQI) Prediction Using Machine Learning

Analyzing and Predicting AQI with RandomForest and XGBoost

Date: April 28, 2025

by Aswin Manohar (47)

Why This Project?



Motive

Air pollution is a global health concern, causing millions of premature deaths annually.

Accurate AQI prediction helps in:

- Monitoring air quality in real-time.
- Informing public health policies and interventions.
- Raising awareness to reduce pollution exposure.

Goal

Develop a reliable model to predict AQI, focusing on high-risk levels, to support better decision-making for environmental and public health.

Project Overview

Objective

Predict AQI values using pollutant data to assess air quality levels.

Dataset

city_day.csv (29,531 rows, 16 columns)

Features: PM2.5, PM10, NO, NO2, NOx, NH3, CO, SO2, O3, Benzene, Toluene, etc.

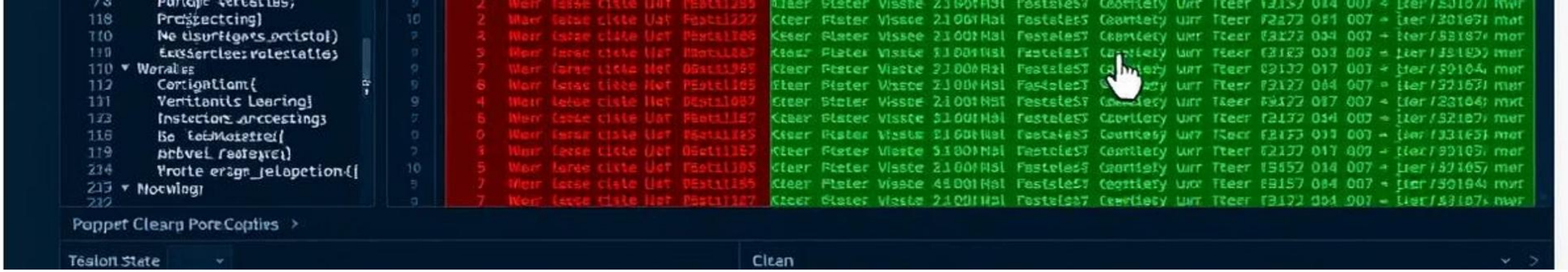
Target: AQI (Air Quality Index)

Approach

Preprocess data (handle missing values, feature selection).

Train models (RandomForest, XGBoost) with hyperparameter tuning.

Improve predictions for higher AQI values.



Data Preprocessing

Initial Steps

Dropped columns with >60% missing values (e.g., Xylene: 61.32% missing).

Filled missing pollutant values with median (e.g., PM2.5, CO).

Filled AQI with forward-fill and dropped rows with remaining null AQI.

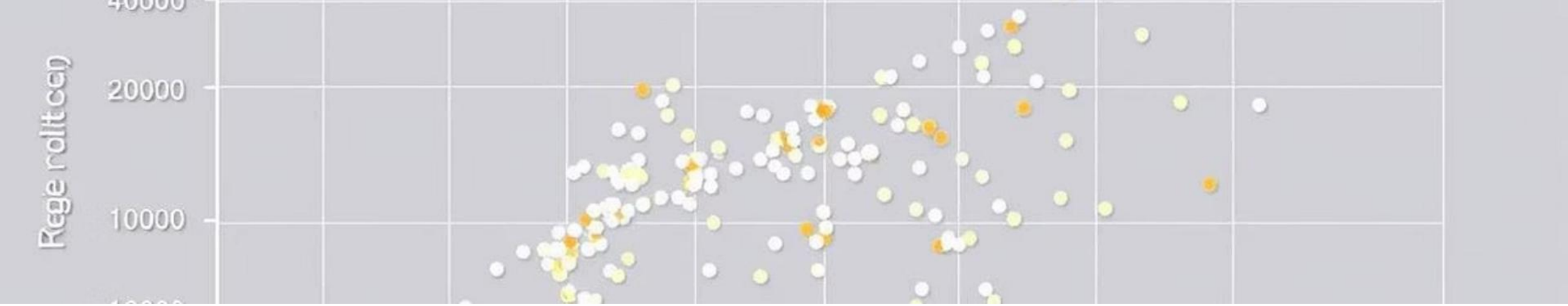
Feature Engineering

Created AQI_Category (Good: ≤50, Moderate: 51-100, Bad: >100).

Final Dataset

Shape: Reduced to 24,850 rows after cleaning.

Key features retained: PM2.5, CO, NO2, Toluene, NO.



Initial Model Performance

Models Used

RandomForest and XGBoost with RandomizedSearchCV for hyperparameter tuning.

Evaluation

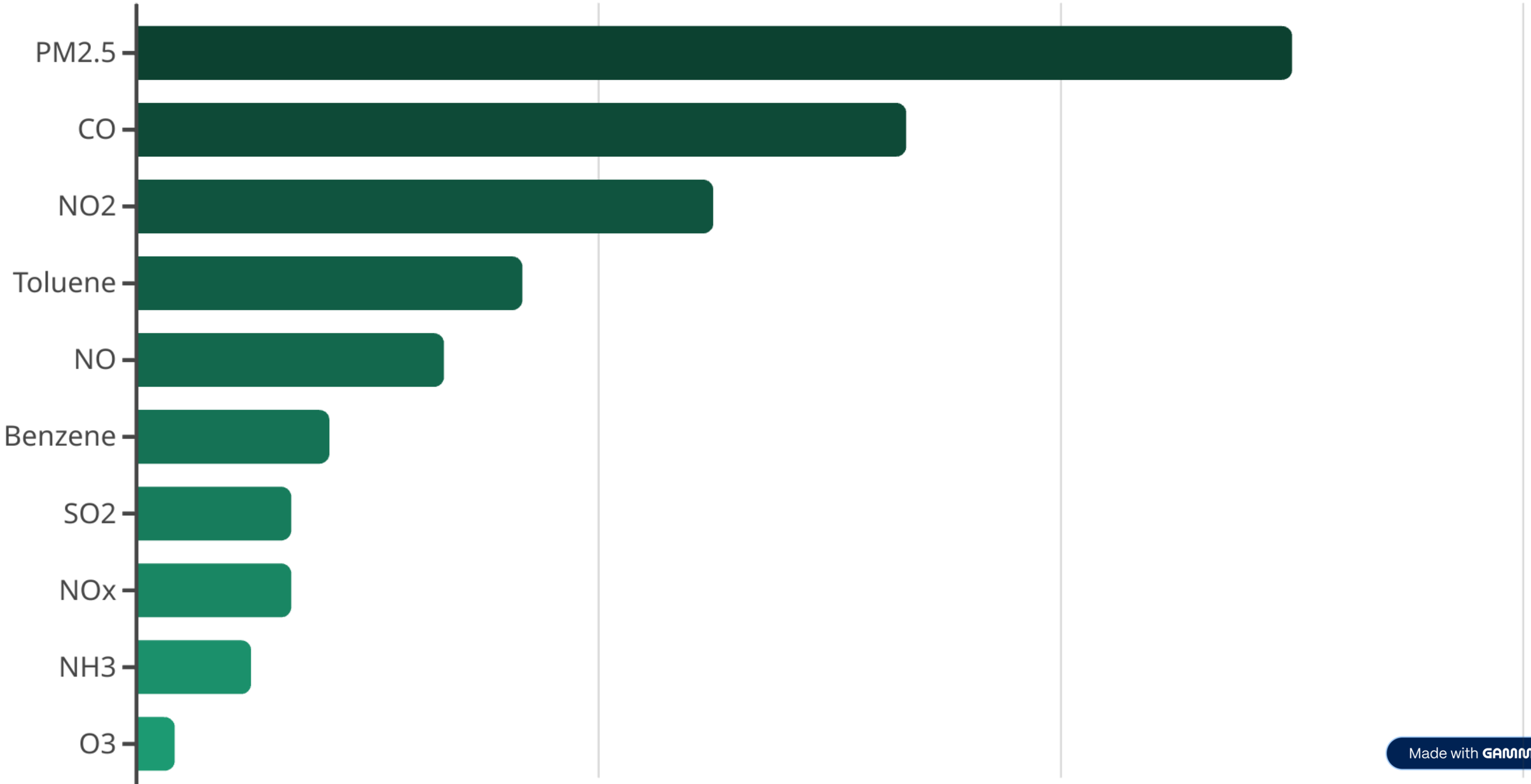
Scatter plot (Predictions vs. Actuals):

- Good accuracy for AQI < 1000.
- Poor performance for AQI > 1000 (high scatter).

Issue Identified

Model struggles with higher AQI values due to data imbalance and feature relevance.

Feature Importance Analysis



Model Improvement Strategies

Custom Loss Function

Implemented weighted MSE in XGBoost to penalize errors on higher AQI values more (weight = 2 for AQI > 1000).

Hyperparameter Tuning

Refined RandomizedSearchCV results with GridSearchCV for better parameters.

Feature Selection

Focused on top features and created interaction terms to capture combined effects.

Updated Model Performance

Actions Taken

Trained XGBoost with top features, interaction terms, and weighted MSE.

Evaluation

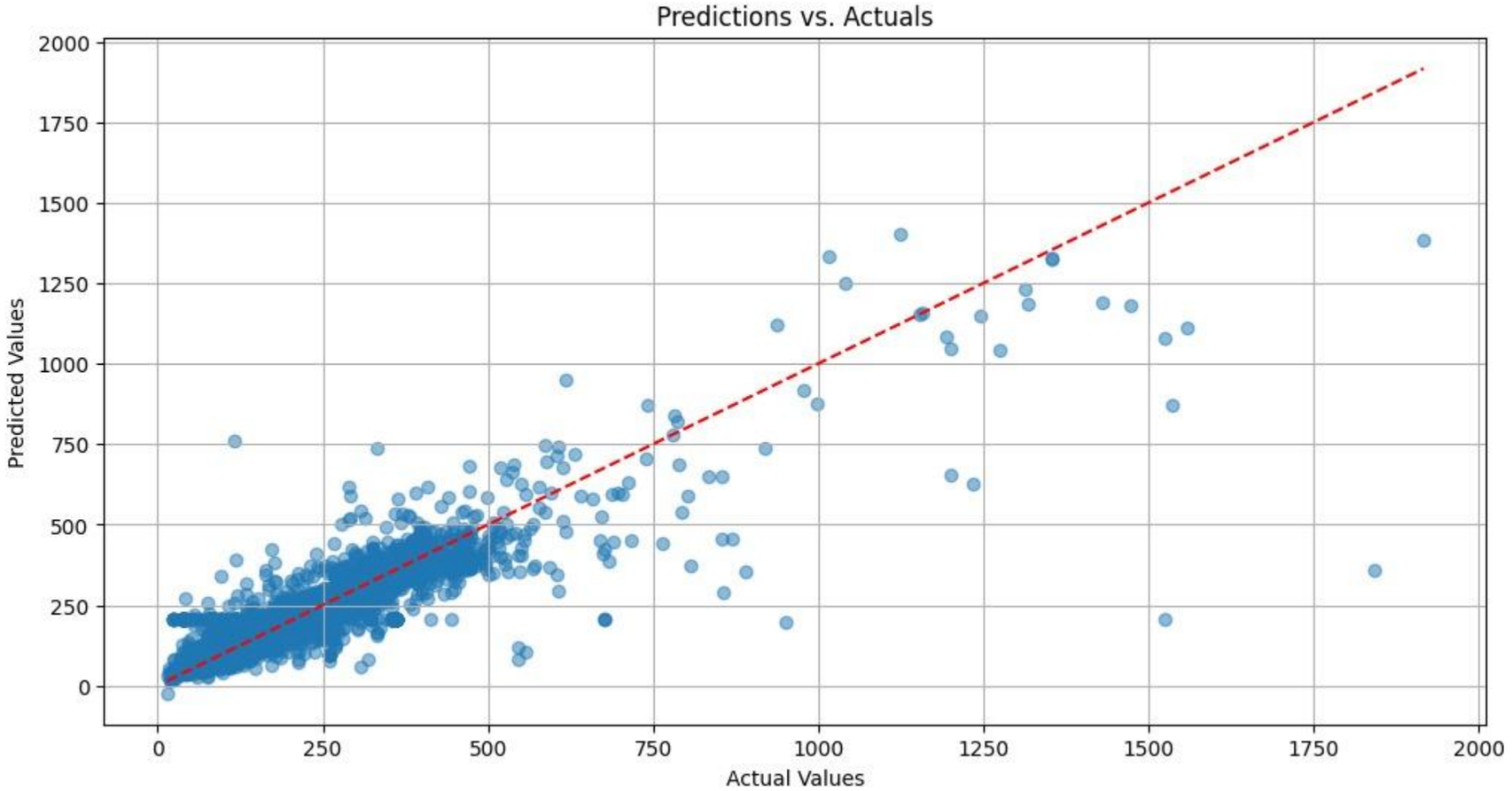
Replot Predictions vs. Actuals (pending new scatter plot).

Calculate MSE and R-squared, focusing on AQI > 1000.

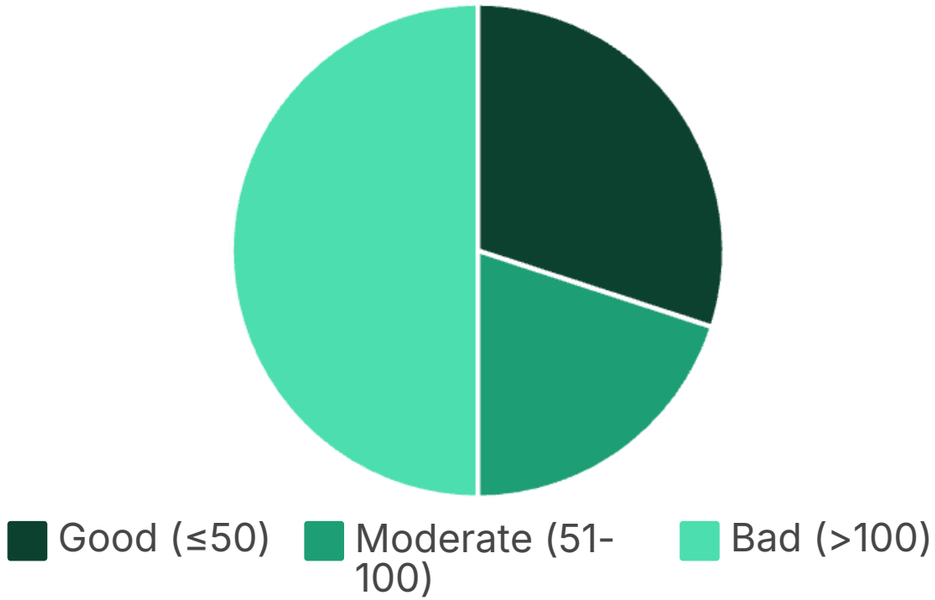
Expected Improvement

Reduced scatter for higher AQI values.

Better alignment along the ideal prediction line.



AQI Category Distribution for Classification



Majority of data points are in the "Bad" category, indicating a need to focus on higher AQI predictions.

Challenges and Solutions

- High missing values in the dataset (e.g., PM10: 37.72% missing).
- Poor prediction accuracy for high AQI values.
- Imbalanced data distribution (more low AQI values).

Next Steps

- Evaluate updated model with new metrics.
- Collect more data for high AQI values.
- Test ensemble methods (stacking RandomForest and XGBoost).
- Incorporate time-series analysis.
- Deploy model for real-time AQI prediction.

Classification vs Regression for AQI

1. **Classification – Predict AQI_Category (e.g., Good, Moderate, Bad)**
   - **Accuracy:** 88% — quite good overall. Weighted Avg F1-Score: 0.87 — solid, especially given class imbalance.
   - **Weighted Avg F1-Score**: 0.81 — indicates the model does reasonably well across all classes, though there's room for improvement,
   - especially for the "Good" class.
   - Target: Categorical (e.g., labels like "Good", "Moderate", "Poor")
   - Evaluation Metrics: Accuracy, Precision, Recall, F1-score, Confusion Matrix
   - Pros:
     - Interpretable output (e.g., "This area is Bad")
     - Easier for alert systems or apps (e.g., show a red warning icon)
   - Cons:
     - Less precise — doesn’t tell you if AQI is 151 or 249, just that it’s "Unhealthy“.
     - Sometimes struggles with class imbalance

Classification vs Regression for AQI

2. Regression – Predict exact AQI value

- Target: Continuous (e.g., AQI = 215.6)
- Evaluation Metrics: RMSE, MAE, R^2 Score
- Result:
  - RMSE: ~63 → decent if AQI values range widely
  - R^2 Score: 0.805 → strong correlation; model explains 80.5% of variance
- Pros:
  - Gives exact AQI — useful for scientific, health, and environmental apps
  - Can be converted to categories post-prediction (e.g., using AQI breakpoints)
- Cons:
  - Less interpretable for general users
  - Sensitive to outliers/extreme AQI values

GitHub

Aswin Manohar
https://github.com/Aswin12408600/Air-Quality-Index-Prediction.git