Predicting Air Quality Using Machine Learning

A CLASSIFICATION APPROACH WITH DECISION TREES AND RANDOM FOREST

BRIAN KIPYEGON

5TH JUNE 2025

Business Problem

- Air pollution impacts health and the environment globally.
- Goal: Classify air quality as Good, Moderate, Poor, or Hazardous.
- Target metric: Achieve > 0.8 recall for reliable predictions.

Project Objectives

- 1. Clean and preprocess the dataset.
- 2. Analyze feature relationships.
- 3. Train and evaluate classification models.
- 4. Select the best model and derive actionable insights.

Dataset Overview

- 5,000 observations across key variables:
- Pollutants: CO, PM2.5, PM10, NO_2, SO_2
- Weather: Temperature, Humidity
- Demographics: Population Density, Industrial Proximity
- Target: Air Quality Level

Tools and Libraries

- Python: Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn
- Models: Logistic Regression, Decision Trees, Random Forest, XGBoost
- Jupyter Notebook

Data Preprocessing

- Handled outliers.
- Label Encoded the Multi-class Target variable.
- Scaled features for consistent input across models.

Feature Importance

- Top Predictors:
- CO (most influential)
- NO_2 and SO_2 (moderate)
- Demographic factors like population density were also influential.
- PM2.5, PM10 and Humidity were less influential and thus weren’t used in modelling

Baseline Model – Decision Trees

- Best Params: max_depth=5, min_samples_leaf=5, etc.
- Recall (Train/Test): 0.845 / 0.846
- Generalization: Excellent (No overfitting)

Random Forest Model

- Best Params: 150 trees, max_depth=8, max_samples=0.6, etc.
- Recall (Train/Test): 0.845 / 0.846
- Generalization: Excellent
- Slightly better performance due to ensemble nature

Model Comparison

METRIC	DECISION TREE	RANDOM FOREST
Train Recall	0.845	0.845
Test Recall	0.846	0.846
Overfitting	No	No

Model Selection

Performance:

- Both Decision Tree and Random Forest achieved identical recall scores, making them both viable from an accuracy standpoint.

Complexity vs. Interpretability:

- *Random Forest*: More accurate on large datasets but computationally expensive and less interpretable.
- *Decision Tree*: Simpler, easier to interpret, and resource-efficient.

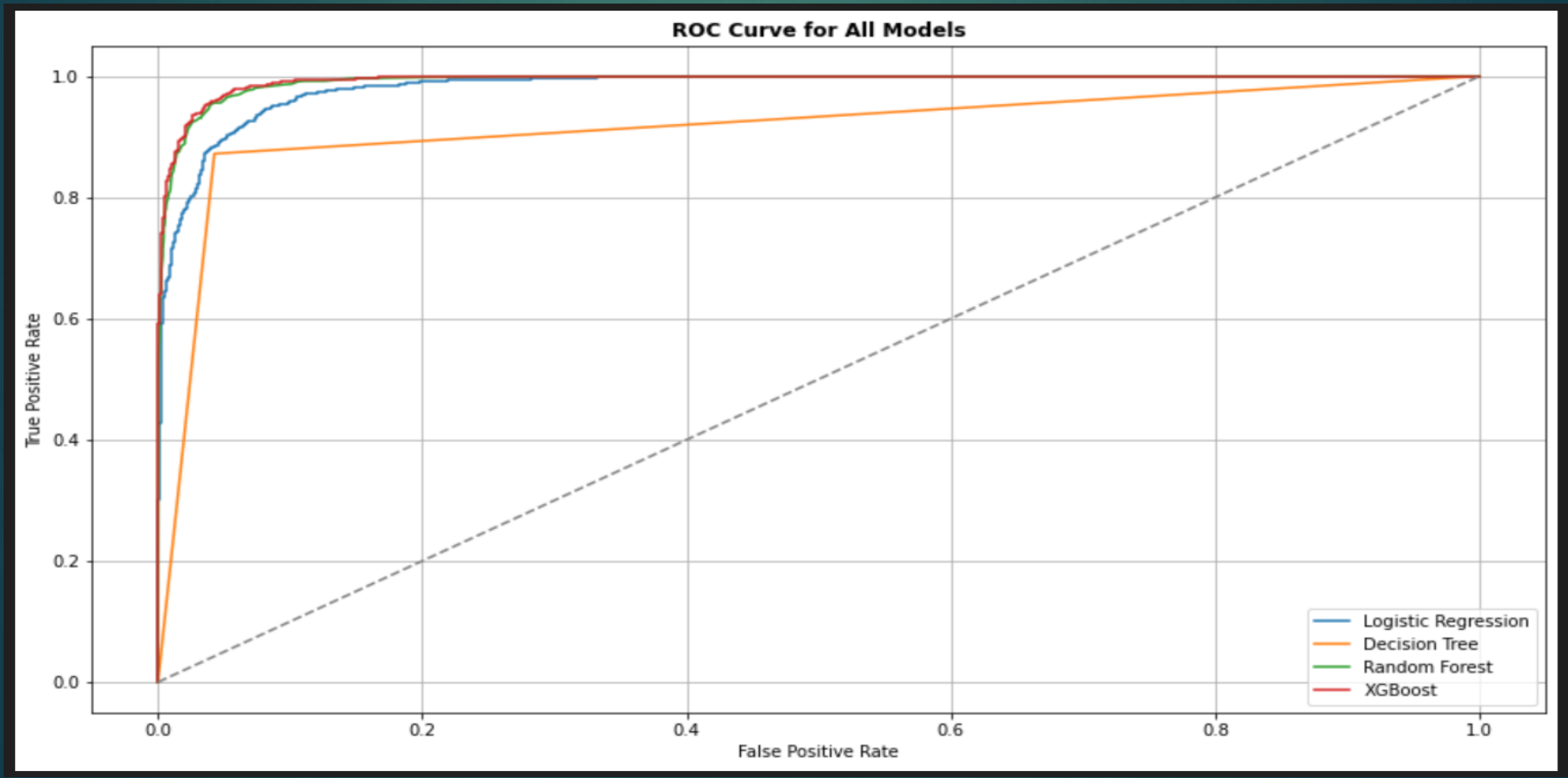
Recommendation:

- Decision Trees Selected due to its efficiency, interpretability, and real-time suitability.

Visualizations

ROC Curve for Multiple Models

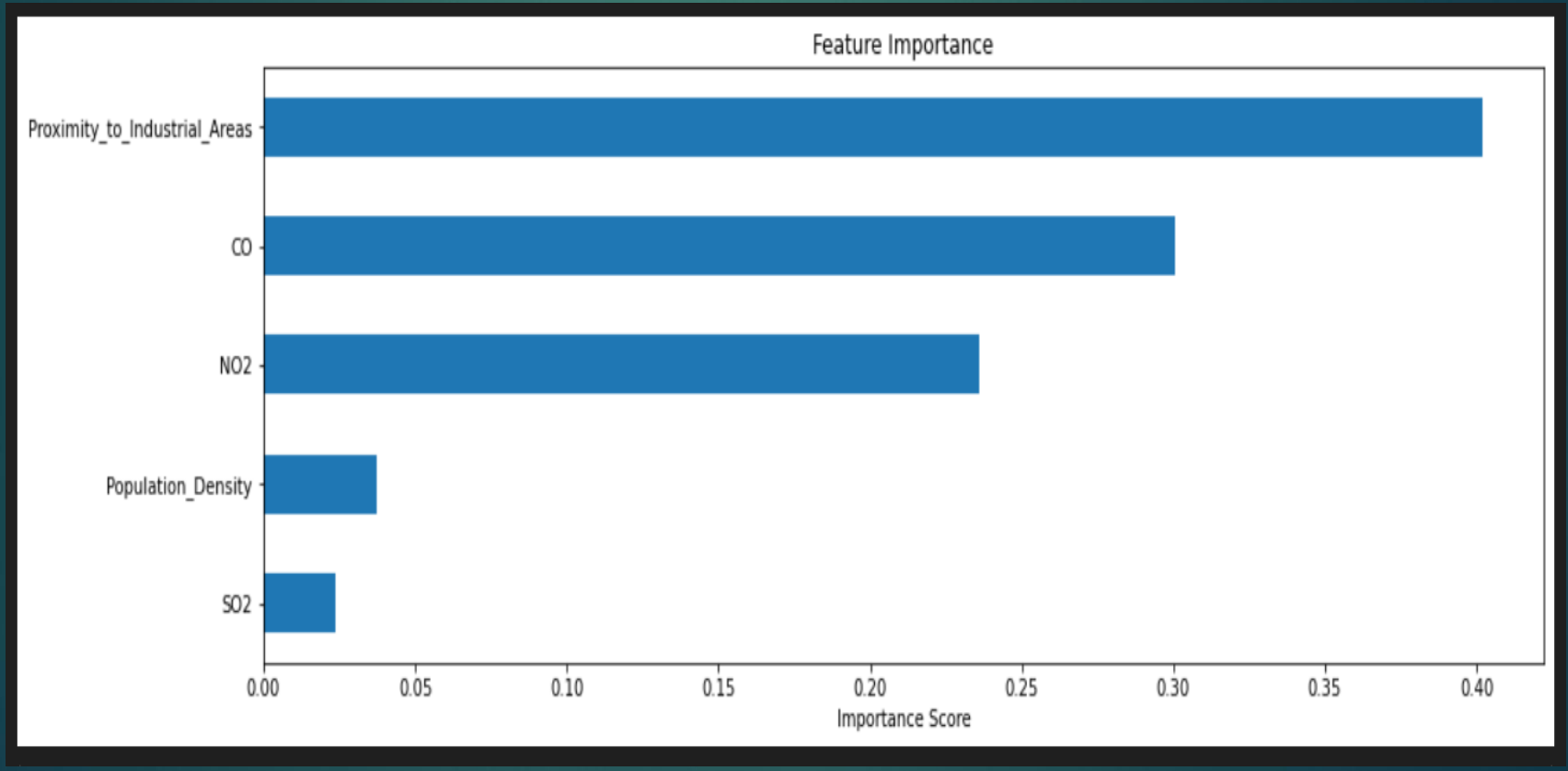
- XGBoost and Random Forest show the highest classification performance with near-perfect ROC curves, indicating strong model reliability. Decision Tree lags behind.



Visualizations

Feature Importance Plot

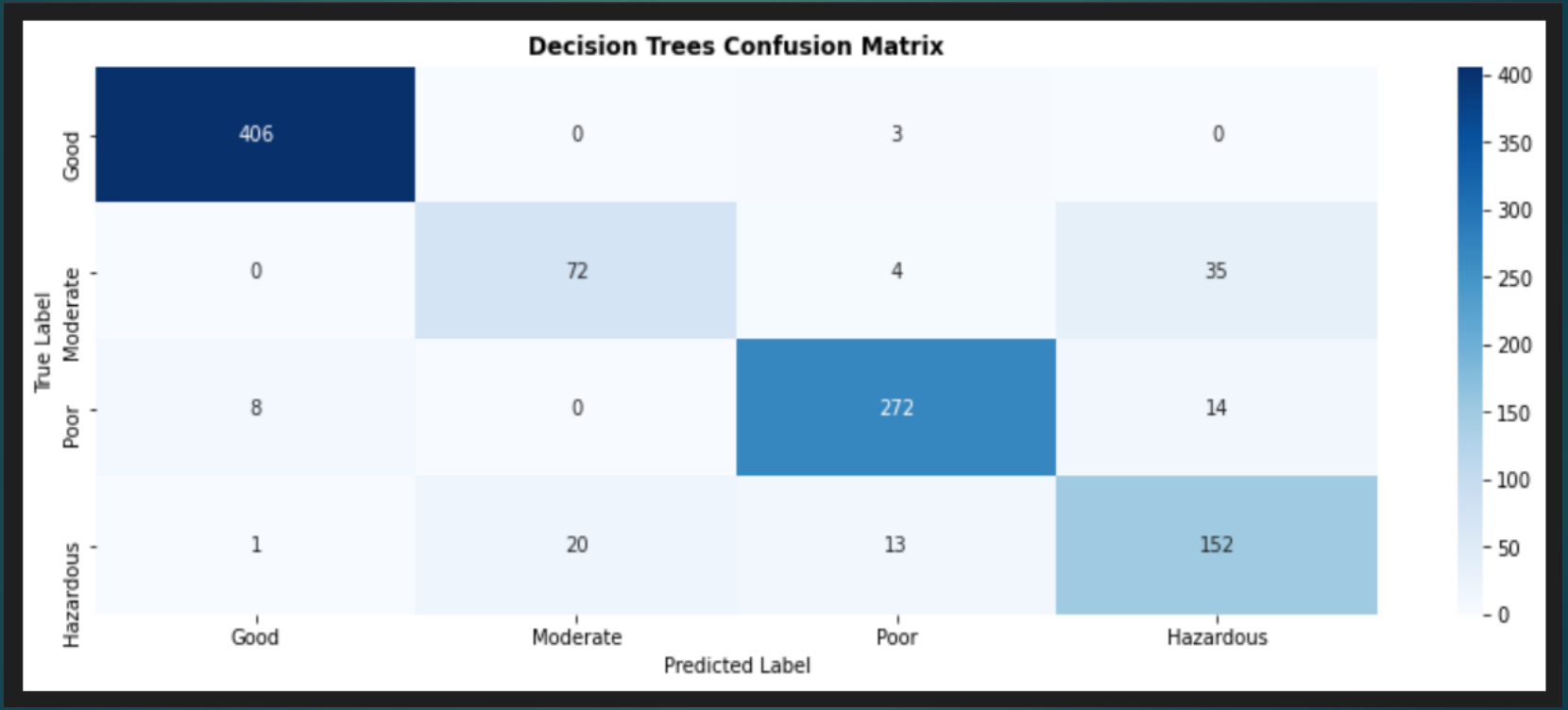
- Proximity to industrial areas is the strongest predictor of quality of air closely followed by CO,NO2,Population_Density and lastly SO2.



Visualizations

Confusion Matrix

- The model performs best on the 'Good' and 'Poor' classes, with high accuracy in predictions.
- However, it struggles more with ‘Moderate’ and especially ‘Hazardous’ cases, showing higher misclassification rates in these classes.



Recommendations

- Focus on monitoring CO, NO_2, and SO_2.
- Use trained Decision Trees model for predictions.
- Guide environmental health policies using model insights.

Conclusion

- Successfully built and validated a reliable classifier.
- Achieved target recall score > 0.8.
- Model is ready for deployment or integration with air monitoring systems.