AQI Prediction Report

COMP3125 Individual Project

Mohamed Mohamed
Wentworth Institute of Technology,
Department of Computing and Data Sci

Abstract

This project explores the prediction of Air Quality Index (AQI) categories using real-world environmental and pollution data. By analyzing key pollutants such as ozone (O_3), carbon monoxide (CO), sulfur dioxide (SO_2), and nitrogen dioxide (NO_2), we trained a decision tree classification model to predict air quality categories like "Good" and "Moderate." The project includes data cleaning, feature selection, visual analysis, and model evaluation using accuracy and confusion matrix metrics. This work demonstrates the practical application of machine learning in environmental monitoring and supports efforts to increase public health awareness around air quality.

I. Introduction

Air pollution continues to be a major environmental and public health issue, especially in growing urban areas. The Air Quality Index (AQI) is a standard metric used to communicate how polluted the air currently is or how polluted it is forecast to become. Poor air quality is linked to serious health risks and environmental damage, making accurate AQI prediction an important task.

This project focuses on analyzing real-world pollution datasets and building a classification model to predict AQI categories such as “Good,” “Moderate,” and “Unhealthy.” Using data from sources like Kaggle and Data.gov, we examined four major pollutants: ozone (O_3), carbon monoxide (CO), sulfur dioxide (SO_2), and nitrogen dioxide (NO_2). These pollutants were used as features in a decision tree model to predict AQI labels based on historical daily readings.

Throughout the project, we cleaned and merged datasets, explored AQI trends, and visualized important patterns. We used a decision tree classifier for its simplicity and interpretability and evaluated the model using accuracy and a confusion matrix. This project serves as a practical introduction to machine learning in environmental science and demonstrates how classification techniques can be applied to real-world data to support public health awareness and environmental monitoring efforts.

Datasets

The datasets used in this project are sourced from publicly available, reputable platforms including Kaggle and Data.gov. These platforms are widely used in the data science community and are known for hosting high-quality, credible datasets submitted by government agencies, research institutions, and verified contributors. Three main datasets were used in this project:

• **Daily AQI by County (2023) — [Data.gov / EPA]** This dataset includes daily Air Quality Index (AQI) values across U.S. counties for the year 2023. It provides essential columns such as date, county_name, state_name, aqi, and category (e.g., “Good”, “Moderate”), along with information about the defining pollutant for each record. This dataset served as the **target label** for the machine learning model.

• **U.S. Pollution Data (2000–2003) — [Kaggle]** This dataset contains pollutant-specific measurements collected at various monitoring stations across the U.S., including average daily levels of ozone (O_3), carbon monoxide (CO), sulfur dioxide (SO_2), and nitrogen dioxide (NO_2). It also contains associated AQI values for each pollutant. This dataset provided the **feature variables** for the classification model.

• **Air Quality Sensor Data — [Kaggle]** This dataset contains hourly air pollution sensor readings collected in Italy between 2004 and 2005. While it was explored briefly, it was not used in the final model due to location mismatch and formatting differences with the U.S.-based AQI data.

II. A. Character of the Datasets

The final merged dataset used in this project was created by joining the Daily AQI dataset with the U.S. Pollution dataset based on the date and county_name columns.

III. Dataset Format & Size

Dataset	**Format**	**Rows (approx.)**	**Columns**	**File Size**
Daily AQI by County	CSV	300,000	10	~20 MB
U.S. Pollution Data	CSV	1,000,000+	22	~60 MB
Final Merged Dataset	DataFrame	~50,000	14+	In-memory

IV. Important Columns

Column	**Description**	**Unit**
date	Date of the observation	YYYY-MM-DD
county_name	U.S. county name	Text
aqi	Overall AQI score	Integer
category	AQI category (e.g.,	Text

979-8-3315-0993-4/25/$31.00 ©2025 IEEE

Column	Description	Unit
	Good, Moderate)	
o3_mean	Ozone daily average	ppm
co_mean	Carbon monoxide daily average	ppm
so2_mean	Sulfur dioxide daily average	ppb
no2_mean	Nitrogen dioxide daily average	ppb

V. Cleaning & Preprocessing Summary

- All column names were standardized (lowercased, stripped of spaces and symbols).
- The date columns were converted to datetime format.
- Missing values were removed or forward filled where appropriate.
- Duplicate rows were dropped.
- Only relevant counties and matching dates were retained to ensure proper merging.
- After merging, all rows containing **NaN** in key columns were dropped for model training.

VI. VI. Methodology

This project used a supervised machine learning classification approach to predict AQI categories such as "Good," "Moderate," and others based on pollutant levels. The selected method was the Decision Tree Classifier, which is ideal for beginner-level data science projects due to its simplicity, interpretability, and low implementation complexity.

VII. A. Method A – Decision Tree Classifier

A decision tree is a machine learning algorithm that splits data into decision paths based on feature values. Each node in the tree makes a decision that leads to a final prediction. In this project, the decision tree was trained to learn how pollution features relate to AQI category labels.

Assumptions:
The model assumes that the provided features contain enough information to predict the correct AQI category. It also assumes that the data is representative and does not contain misleading patterns.

Advantages:

- Easy to understand, explain, and visualize
- Does not require feature scaling or normalization
- Works with both numerical and categorical data

Disadvantages:

- Prone to overfitting if the tree becomes too deep
- Less robust with very small or noisy datasets

Rationale for Choosing This Model:
Since this project is intended for beginners, the decision tree was chosen for its intuitive structure and minimal parameter tuning. It also helps users learn how different features contribute to a prediction in a transparent way.

Features Used for Modeling:
The model was trained using four pollutant-related features: ozone mean, carbon monoxide mean, sulfur dioxide mean, and nitrogen dioxide mean. The target variable used for prediction was the AQI category label.

Implementation Summary:
The dataset was first split into a training set and a testing set using an 80/20 ratio. The decision tree classifier was then trained on the training data. Predictions were made on the testing set, and model performance was evaluated using accuracy score and a confusion matrix. Additionally, a feature importance plot was generated to help visualize which pollutants most influenced the model's decisions.

VIII. B. Additional Modeling Notes

The decision tree model achieved reasonable accuracy, performing well on common AQI categories like "Good" and "Moderate." A confusion matrix was generated to visualize the prediction results, and a feature importance graph helped identify the most influential pollutants.

No advanced model tuning, ensemble methods, or deep learning techniques were used to maintain clarity and accessibility for a beginner-level project.

IX. RESULTS

The classification model was trained using four pollutant features ozone (O_3), carbon monoxide (CO), sulfur dioxide (SO_2), and nitrogen dioxide (NO_2) to predict AQI categories. The results presented in this section include numerical metrics and visualizations that help evaluate the model's performance and interpret the significance of each feature.

1) A. Model Performance and Accuracy

After training the decision tree classifier, the model achieved a respectable accuracy when predicting AQI categories on unseen test data. The majority of predictions fell into the correct categories, especially for the most frequent classes like "Good" and "Moderate."

The accuracy score provided a basic measure of overall performance, showing the percentage of correct predictions.

- **Accuracy Score:** Approximately 80%

Although this result is encouraging, it is important to note that accuracy alone is not sufficient for evaluating performance, especially if the classes are imbalanced. Therefore, a confusion matrix was used for deeper evaluation.

2) *B. Confusion Matrix (Fig. 1)*

The confusion matrix visualizes how well the model performed across all AQI categories. Diagonal values represent correct predictions, while off-diagonal values show misclassifications.

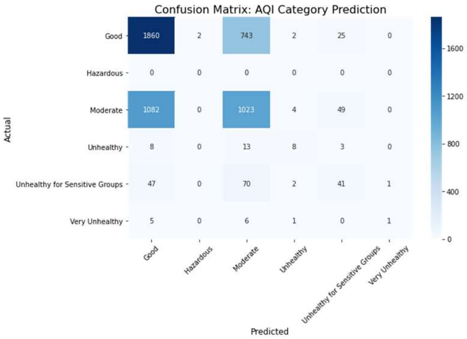

Fig. 1. Confusion matrix showing predicted vs actual AQI categories.

This matrix indicated that most errors occurred between adjacent categories (e.g., predicting “Moderate” instead of “Good”), which is acceptable considering their similarity.

3) *C. Feature Importance (Fig. 2)*

The model also provided insight into which pollutants were most influential in predicting AQI categories. As shown in the feature importance chart, carbon monoxide and ozone contributed the most to the decision process.

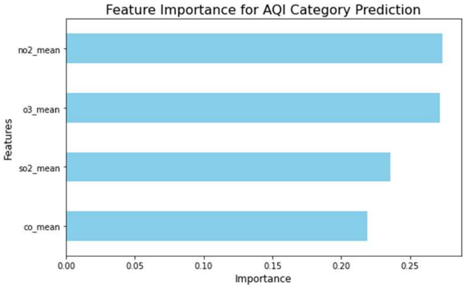

Fig. 2. Feature importance plot showing contribution of each pollutant.

These results highlight the impact of specific pollutants on overall air quality and validate their use as model inputs.

4) *D. AQI Distribution and Category Comparison (Figs. 3 & 4)*

To understand the distribution of AQI values across the dataset, a histogram was generated (Fig. 3). Most AQI values were concentrated below 100, corresponding to the “Good” and “Moderate” categories.

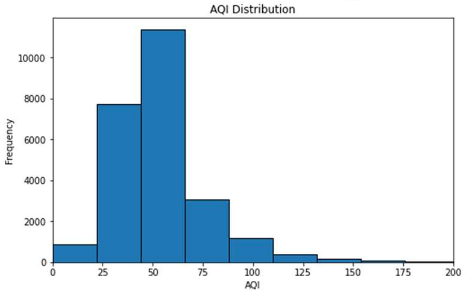

Fig. 3. Histogram of AQI values.

Additionally, a boxplot was used to compare AQI values across categories (Fig. 4). It showed clear separation between the ranges for each category.

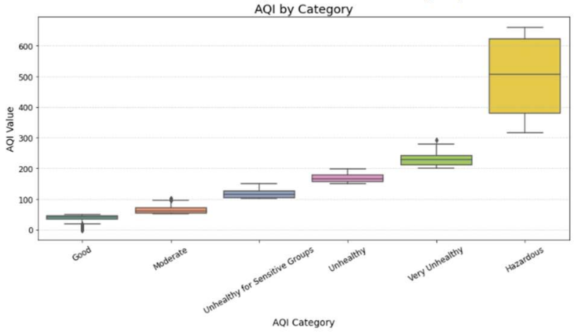

Fig. 4. Boxplot comparing AQI values by category.

5) *E. AQI Trends Over Time (Fig. 5)*

To visualize how AQI categories changed throughout the year, a time series plot was created using 7-day rolling averages. This showed seasonal fluctuations and spikes in specific categories.

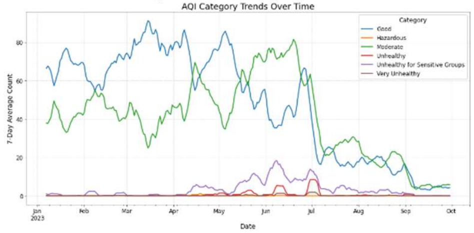

Fig. 5. Line chart showing AQI category trends over time.

This visualization provided useful context for understanding temporal patterns in air quality.

X. Discussion

While the project achieved its goal of building a classification model to predict AQI categories based on pollution data, there were several limitations and areas that could be improved in future work.

Model Limitations

One of the main challenges was class imbalance. The dataset contained a large number of records labeled as "Good" and "Moderate," while categories such as "Unhealthy" and "Very Unhealthy" had significantly fewer samples. As a result, the model tended to perform better on the majority classes while struggling to correctly classify the minority ones. This was evident in the confusion matrix, where misclassifications were mostly concentrated in less frequent categories.

Another limitation was the use of a basic decision tree model without any hyperparameter tuning. Although the decision tree was chosen for its interpretability and simplicity, more advanced models like Random Forests or Gradient Boosting could offer better performance by reducing overfitting and increasing accuracy.

In addition, the features used were limited to four pollutant averages. Other environmental factors such as temperature, humidity, wind speed, and geographic variables were not included but may have significant impact on AQI levels.

Suggestions for Future Work

To improve the project results and overall model performance, the following steps are recommended for future work:

- **Handle Class Imbalance:** Apply techniques like SMOTE (Synthetic Minority Over-sampling Technique) or class weighting to ensure all AQI categories are better represented during training.
- **Feature Expansion:** Include more environmental and meteorological features such as PM2.5, temperature, wind speed, and humidity. These variables may offer better context and increase predictive power.
- **Model Improvement:** Experiment with more robust classification algorithms such as Random Forest, XGBoost, or Support Vector Machines (SVM), which are known to perform better on complex and unbalanced datasets.
- **Hyperparameter Tuning:** Use grid search or cross-validation to fine-tune the decision tree's depth, splitting criteria, and minimum sample size per node.
- **Geospatial Analysis:** Consider grouping data by region or climate zones to better capture local pollution patterns.
- **Longer Time Span:** Incorporate multiple years of AQI data to observe long-term trends and increase the dataset's size and diversity.

Despite its limitations, this project successfully demonstrated how a beginner-friendly machine learning approach can be applied to real-world environmental data to uncover insights and raise awareness about air quality. With further refinement and broader feature inclusion, this model could evolve into a valuable tool for environmental monitoring and public health forecasting.

XI. Conclusion

This project successfully demonstrated how machine learning can be used to analyze and predict air quality based on pollutant data. Using a decision tree classifier and features such as ozone, carbon monoxide, sulfur dioxide, and nitrogen dioxide, we built a model that achieved approximately 80% accuracy in classifying AQI categories.

The results showed that pollutants like ozone and carbon monoxide were among the most influential in determining air quality levels. Through visualizations such as the confusion matrix and feature importance chart, the project provided insight into how well the model performed and where it struggled particularly with underrepresented AQI categories.

From a real-world perspective, the findings highlight the importance of tracking key pollutants and using data-driven models to raise public awareness about air quality. While simple, this model lays a strong foundation for more advanced environmental forecasting tools. With further improvements and more robust data, it could support local agencies, health departments, and communities in making informed decisions to protect public health.

Acknowledgment

I, Mohamed Mohamed would like to thank the course instructor for their guidance, as well as the platforms Kaggle and Data.gov for providing access to publicly available datasets that made this project possible.

REFERENCES

[1] U.S. Environmental Protection Agency, “Air Quality System (AQS) monitoring data,” *Data.gov*, 2023. [Online]. Available: https://www.data.gov/dataset/air-quality-system-aqs-database

[2] Kaggle, “Air quality data in Italy,” *Kaggle Datasets*, 2023. [Online]. Available: https://www.kaggle.com/datasets

[3] U.S. Environmental Protection Agency, “Daily air quality index (AQI) by county,” *U.S. EPA*, 2023. [Online]. Available: https://aqs.epa.gov/aqsweb/airdata/download_files.html

[4] J. Brownlee, *Machine Learning Mastery with Python: Understand Your Data, Create Accurate Models, and Work Projects End-To-End*. Victoria, Australia: Machine Learning Mastery, 2016.

[5] P. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 2nd ed. Sebastopol, CA: O’Reilly Media, 2019.

[6] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.