Air Quality Prediction Report

A Machine Learning-Based Analysis of Air Quality Index

---

1. Introduction

Air quality is a vital indicator of environmental health, significantly impacting human well-being. Understanding the **Air Quality Index (AQI)** and its relation to key pollutants enables governments and stakeholders to manage pollution effectively and improve public health. This report outlines a machine learning approach to predict AQI and classify pollution levels using data from the Iranian Environmental Organization.

AQI	Air Pollution Level	Air Pollution Category	Health Implications	Recommended Precautions
0-50	Level 1	Excellent	No health implications.	Everyone can continue their outdoor activities normally because the air is not polluted.
51-100	Level 2	Good	Some pollutants may slightly affect very few hypersensitive individuals.	Only very few hypersensitive people should reduce outdoor activities.
101-150	Level 3	Lightly Polluted	Healthy people may experience slight irritations and sensitive individuals will be slightly affected to a larger extent because the air is slightly polluted.	Children, seniors and individuals suffering respiratory or heart diseases should reduce sustained and high-intensity outdoor exercises.
151-200	Level 4	Moderately Polluted	Sensitive individuals will experience more serious conditions because the air is moderately polluted. The hearts and respiratory systems of healthy people may be affected.	Children, seniors and individuals with respiratory or heart diseases should avoid sustained and high-intensity outdoor exercises. General population should moderately reduce outdoor activities.
201-500	Level 5	Heavily Polluted	Healthy people will commonly show symptoms. People suffering from respiratory or heart diseases will be seriously affected and will experience reduced endurance in activities.	Children, seniors and individuals with heart or lung diseases should stay indoors and avoid outdoor activities. General population should reduce outdoor activities.

---

2. Dataset Overview

The dataset consists of:

- **2123 air pollution monitoring stations**
- Metrics included:
    - **CO** (Carbon Monoxide)
    - **O3** (Ozone)
    - **NO2** (Nitrogen Dioxide)
    - **SO2** (Sulfur Dioxide)
    - **PM10** (Particulate Matter 10)
    - **PM2.5** (Particulate Matter 2.5)

There is AQI as target.

Data Preprocessing Steps:

- **Missing Values:** Resolved using the median.

- **Duplicated Rows:** Removed.
- **Column Names and Contents:** Originally written in Persian; converted to English.

Two target variables were created:

1. **AQI Average**
2. **AQI Description**

---

3. Regression Analysis

3.1 Models Used

To predict the **AQI Average**, the following regression models were evaluated:

- Gradient Boosting
- XGBoost
- Random Forest
- KNN
- Decision Tree
- SVR
- Linear Regression, Ridge, Lasso, ElasticNet

3.2 Evaluation Metrics

- Mean Absolute Error (MAE)
- R^2 (Coefficient of Determination)

3.3 Results

The **best models** were:

Rank	Model	MAE	R^2
1	Gradient Boosting	4.576557	0.909186
2	Random Forest	4.947887	0.905484
3	XGBoost	4.962615	0.905852
4	KNN	5.127256	0.873260
5	Decision Tree	5.332536	0.857144
6	SVR	6.478604	0.813697
7	ElasticNet	13.003160	0.693486
8	Lasso	13.003974	0.693382
9	Linear Regression	13.018671	0.691836
10	Ridge	13.018671	0.691836

Best Hyperparameters:

Gradient Boosting :

- learning_rate : 0.05
- max_depth : 10
- min_samples_split : 10
- n_estimators : 300
- subsample : 0.8

3.4 Feature Importance

For the **Gradient Boosting** model, feature importance was as follows:

Feature	Importance
PM2.5	0.697276
CO	0.105519
NO2	0.063716
O3	0.053219
PM10	0.053100
SO2	0.027170

Observation:
PM2.5 has the highest importance, indicating its dominant contribution to AQI prediction.

---

4. Classification Analysis (AQI Description)

4.1 Models Used

For classifying AQI into **"Excellent," "Good," "Lightly Polluted," "Moderately Polluted," and "Heavily Polluted"**, the following classifiers were implemented:

- KNN
- XGBoost
- Gradient Boosting
- Random Forest
- Decision Tree
- SVC
- Logistic Regression

4.2 Evaluation Metrics

- Accurracy
- Recall
- F1-Score
- ROC-AUC

4.3 Results

The **best models** was:

Model	Accuracy	ROC AUC	F1-score	Recall
Gradient Boosting	0.9426	0.9840	0.7251	0.7222
XGBoost	0.9330	0.9715	0.7285	0.7266
Random Forest	0.9282	0.9897	0.7219	0.7103
KNN	0.9258	0.9444	0.9268	0.9131
Decision Tree	0.9067	0.8367	0.7081	0.7029
SVC	0.8612	0.9707	0.6777	0.6713
Logistic Regression	0.7727	0.8662	0.6108	0.5854

Best Hyperparameters:

Gradient Boosting :

- learning_rate: 0.1
- max_depth: 7
- n_estimators: 200
- subsample: 0.8

4.4 Feature Importance

Feature	Importance
PM2.5	0.596267
NO2	0.112122
CO	0.105613
PM10	0.087409
O3	0.055173
SO2	0.043416

---

5. Deep Learning (Multilayer Perceptron)

5.1- MLP model_Classification:

5.1.1-Best Hyperparameters:

- batch_size=16,
- dropout_rate=0.4
- epochs=550
- learning_rate=0.01,
- neurons_layer1=32
- neurons_layer2=16

5.1.2- Feature Importance:

PM2.5	0.5308
PM10	0.1554
NO2	0.1149
CO	0.0918
O3	0.0604
SO2	0.0467

5.2- MLP model_Regression:

5.2.1-Best Hyperparameters:

- batch_size: 22
- epochs: 381
- model__dropout_rate: 0.11060969310248417
- model__learning_rate: 0.0018505523167281056
- model__neurons: 122

5.2.2- Feature Importance:

PM2.5	0.6051
CO	0.1024
PM10	0.0872
NO2	0.0818
O3	0.0711
SO2	0.0524

---

6. Feature Importance and Analysis

6-1-Regression Analysis: Gradient Boosting and MLP Models

1. **Gradient Boosting Regression:**
    - **PM2.5** is the most influential feature with a feature importance of **0.697**.
    - Other significant contributors include **CO (0.106)**, **NO2 (0.064)**, **O3 (0.053)**, **PM10 (0.053)**, and **SO2 (0.027)**.
    - **Observation:** The dominance of PM2.5 highlights its critical impact on AQI prediction, likely due to its direct and severe health implications.
2. **MLP Regression:**
    - **PM2.5** holds the highest normalized permutation importance of **0.6051**, followed by **CO (0.1024)**, **PM10 (0.0872)**, **NO2 (0.0818)**, **O3 (0.0711)**, and **SO2 (0.0524)**.
    - **Observation:** Both models agree on the importance of PM2.5, but slight variations in other features suggest differences in model learning.

6-2-Classification Analysis: Gradient Boosting and MLP Models

1. **Gradient Boosting Classification:**
    - **PM2.5** is again the leading feature with an importance score of **0.5963**.
    - Secondary contributors include **NO2 (0.1121)**, **CO (0.1056)**, **PM10 (0.0874)**, **O3 (0.0552)**, and **SO2 (0.0434)**.

- **Observation:** Consistent importance of PM2.5 emphasizes its critical role in determining pollution levels across categories.

2. **MLP Classification:**
    - **PM2.5** leads with a feature importance of **0.5308**, followed by **PM10 (0.1554)**, **NO2 (0.1149)**, **CO (0.0918)**, **O3 (0.0604)**, and **SO2 (0.0467)**.
    - **Observation:** The MLP model assigns slightly higher importance to PM10 compared to Gradient Boosting, possibly capturing non-linear relationships.

---

7. Conclusion

Dominance of PM2.5: Across all models, PM2.5 consistently emerges as the most critical feature, emphasizing its direct impact on air quality and human health.

Consistency Across Models: Gradient Boosting and MLP models align in identifying PM2.5, CO, NO2, PM10, O3, and SO2 as key contributors, reflecting the robustness of feature importance analysis.

Model Performance:

- Gradient Boosting outperformed others in both regression (MAE: **4.576557**) and classification (Accuracy: **0.9426**), highlighting its suitability for AQI tasks.
- MLP models demonstrated competitive performance with flexibility in feature importance.

---

8. Recommendations and Future Work

Target PM2.5 in Interventions:

- Focus regulatory measures and public health campaigns on reducing PM2.5 levels, as it is the most impactful feature across all models.

Feature Engineering:

- Incorporate meteorological factors such as temperature, humidity, and wind speed to improve predictions. These factors often influence pollutant dispersion and concentration.

Regionalized Models:

- Develop separate models for specific regions to account for localized factors like traffic density, industrial activities, and seasonal trends.

Model Combination:

- Use ensemble approaches (e.g., stacking Gradient Boosting and MLP models) to leverage the strengths of both models for robust predictio

---