

## GB Air Quality Forecasting with Machine Learning (PM2.5 Prediction)

*A Multi-City Air Pollution Analysis & Forecasting Project using API Integration, Feature Engineering, Random Forest, and XGBoost*

---

### Overview

This project aims to **predict PM2.5 air pollution levels** using real-time and historical environmental measurements collected from **5 major cities in Türkiye** via the OpenWeatherMap API.

The dataset includes:

- Air pollution components (CO, NO, NO2, O3, SO2, NH3, PM10)
- Meteorological data (temperature, humidity, wind speed)
- Time-based features (hour, weekday, month)
- City encoding
- Target variable: **PM2.5**

A full ML pipeline is implemented including:

- ✓ API data collection
- ✓ Time-series feature extraction
- ✓ Data preprocessing (scaling + encoding)
- ✓ Model training
- ✓ Model evaluation
- ✓ Visualization (feature importance + prediction vs actual)

Two machine learning models are compared:

- **Random Forest Regressor**
  - **XGBoost Regressor**
- 

### Cities Covered (5-city dataset)

Data was collected for:

- **Istanbul**
- **Ankara**
- **Izmir**
- **Bursa**
- **Antalya**

Each city contributes 5 days of historical air pollution data + real-time weather data.

---

## Project Pipeline

### **1 Data Collection (API Integration)**

Using OpenWeatherMap Air Pollution and Weather APIs:

```
BASE_URL = "http://api.openweathermap.org/data/2.5/air_pollution"
```

```
WEATHER_URL = "http://api.openweathermap.org/data/2.5/weather"
```

Collected features include:

- AQI
  - PM2.5 (target)
  - PM10
  - Gas pollutants (CO, NO, NO2, O3, SO2, NH3)
  - Temperature
  - Humidity
  - -City code
  - Time metadata
- 

### **2 Feature Engineering**

Additional features are created:

```
df["hour"]
```

```
df["weekday"]
```

```
df["month"]
```

These improve temporal prediction patterns.

---

### **3 Data Preprocessing**

- Numeric features scaled using **StandardScaler**
- City names encoded using **LabelEncoder**
- Missing values automatically filled using column means

Outputs saved as:

```
extended_air_data.csv
```

```
processed_extended.csv
```

---

### **4 Machine Learning Models**

Two regression algorithms were trained:

#### ✓ Random Forest

- Ensemble of decision trees
- Stable baseline model
- Feature importance available

#### ✓ XGBoost

- Advanced gradient boosting
  - Higher accuracy
  - Fast and robust for tabular data
- 

### 5 Model Evaluation

Metrics calculated:

- **RMSE** – Root Mean Squared Error
- **MAE** – Mean Absolute Error
- **R<sup>2</sup> Score** – Explained variance

*(Actual values not provided here; left intentionally blank for general documentation.)*

---

### Visual Results

#### ◆ Random Forest – Feature Importance

Shows PM10, NH3, CO, SO2 as strongest predictors.

#### ◆ Random Forest – Actual vs Predicted

Strong linear relationship visible.

---

#### ◆ XGBoost – Feature Importance

XGBoost gives more weight to:

- PM10
- Encoded city
- NH3
- Temperature & humidity

#### ◆ XGBoost – Actual vs Predicted

Tighter predictions around the diagonal → better accuracy.

---

## Repository Structure

air-quality-forecast/

- |— air\_quality\_final\_project.ipynb
- |— extended\_air\_data.csv
- |— processed\_extended.csv
- |— Random\_Forest\_feature\_importance.png
- |— Random\_Forest\_prediction\_vs\_actual.png
- |— XGBoost\_feature\_importance.png
- |— XGBoost\_prediction\_vs\_actual.png
- |— README.md

---

## Tech Stack

- Python
- Pandas, NumPy
- Scikit-Learn
- XGBoost
- Matplotlib & Seaborn
- OpenWeatherMap API
- Jupyter Notebook

---

## How to Run

pip install pandas numpy scikit-learn xgboost matplotlib seaborn requests

python YOUR\_FILE\_NAME.py

---

## Conclusion

- Both models show strong performance in predicting PM2.5
- **XGBoost** tends to perform better with higher accuracy
- PM10, NH3, CO, and SO2 are strong indicators of PM2.5
- Time-based features also improve prediction stability

This project demonstrates:

- ✓ API integration
- ✓ Automated data collection
- ✓ End-to-end ML pipeline
- ✓ Environmental data modeling

---

## Developer

**Busenur Durak**

GitHub: <https://github.com/busenur-durak>

LinkedIn: <https://linkedin.com/in/busenur-durak>

---

## Hava Kalitesi Tahmini (PM2.5 Öngörüsü) – Makine Öğrenimi Projesi

---

### Proje Özeti

Bu proje, Türkiye'nin 5 büyük şehrinden toplanan hava kalitesi ve meteorolojik verileri kullanarak **PM2.5 değerlerini tahmin etmek** için geliştirilmiştir.

Veriler OpenWeatherMap API üzerinden toplanmış ve uçtan uca bir makine öğrenimi pipeline'ı uygulanmıştır:

- ✓ API veri toplama
  - ✓ Zaman bazlı özellik mühendisliği
  - ✓ Veri ölçeklendirme ve encoding
  - ✓ Random Forest ve XGBoost model eğitimi
  - ✓ Hata metrikleri
  - ✓ Görselleştirme
- 

### Kapsanan Şehirler (5 şehir)

- İstanbul
  - Ankara
  - İzmir
  - Bursa
  - Antalya
- 

### Özellikler

Veri seti şu bilgileri içerir:

- Gaz kirliliği bileşenleri (CO, NO2, O3, SO2, NH3, PM10)

- PM2.5 (hedef değışken)
- Sıcaklık
- Nem
- Rüzgar hızı
- Saat / gün / ay
- Şehir kodu

---

## Makine Öğrenimi Modelleri

### ✓ Random Forest

Kararlı bir ensemble modelidir.

### ✓ XGBoost

Daha yüksek doğruluk göstermiştir (boosting tabanlı).

---

## Görsel Sonuçlar

- Özellik önemi grafiklerinde PM10 en güçlü değışken olarak görülüyor.
- XGBoost modelleri hedef değeri daha iyi yakalıyor.
- Gerçek vs Tahmin grafikleri güçlü doğrusal ilişki gösteriyor.

---

## Dosya Yapısı

air-quality-forecast/

|— air\_quality\_final\_project.ipynb  
|— extended\_air\_data.csv  
|— processed\_extended.csv  
|— Random\_Forest\_feature\_importance.png  
|— Random\_Forest\_prediction\_vs\_actual.png  
|— XGBoost\_feature\_importance.png  
|— XGBoost\_prediction\_vs\_actual.png  
|— README.md

---

## Teknolojiler

- Python

- Pandas, NumPy
  - Scikit-Learn
  - XGBoost
  - Matplotlib, Seaborn
  - OpenWeatherMap API
- 



## Sonuç

Bu projede:

- PM2.5 başarılı şekilde tahmin edilmiştir
  - En iyi performansı XGBoost göstermiştir
  - PM10, NH3, CO ve SO2 en etkili değişkenlerdir
  - Zaman temelli özellikler tahminleri güçlendirmiştir
- 



## Geliştirici

**Busenur Durak**

GitHub: [github.com/busenur-durak](https://github.com/busenur-durak)

LinkedIn: [linkedin.com/in/busenur-durak](https://linkedin.com/in/busenur-durak)