VARNA AIR

PROJECT PROPOSAL

28.06.2025

IVET KALCHEVA

CONTENTS

INTRODUCTION

WHAT

This project aims to predict the Air Quality Index (AQI) for Varna, Bulgaria, using regression models. The primary focus is on leveraging environmental data such as pollutant levels (CO, NO, NO_2, O_3, SO_2, PM2_5, PM10 and NH_3) to build a predictive model.

WHY

Air quality is a significant public health issue, impacting millions of lives globally. For me, this project is deeply personal because Varna is my hometown, where my family and loved ones live. By predicting AQI for Varna, I aim to contribute to the well-being of the community and provide actionable insights for improving urban living conditions.

WHO

The primary stakeholders are environmental agencies, urban planners and public health organizations in Varna. Additionally, my mother, a Varna resident, represents individuals who are directly impacted by air quality changes and can provide valuable insights into the lived experiences of the population.

WHEN

This project will be completed over 2 weeks during the Open Programme.

The **first week** will focus on acquiring, cleaning and exploring the dataset, ensuring that all relevant environmental factors are accurately represented and any inconsistencies addressed.

The **second week** will be dedicated to model development, training and evaluation, alongside the creation of detailed visualizations and actionable insights.

HOW

By acquiring authentic and diverse datasets, cleaning and analysing them, and applying regression models, I will predict AQI values and evaluate the model's performance using metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and R^2.

DOMAIN UNDERSTANDING

RESEARCH QUESTION

How can environmental factors be used to accurately predict the Air Quality Index (AQI)?

EXPLORATORY RESEARCH

- Which pollutants (CO, NO, NO_2, O_3, SO_2, PM2_5, PM10 and NH_3) have the strongest correlation with AQI values?
- Are there any temporal trends or seasonal variations in pollutant concentrations?

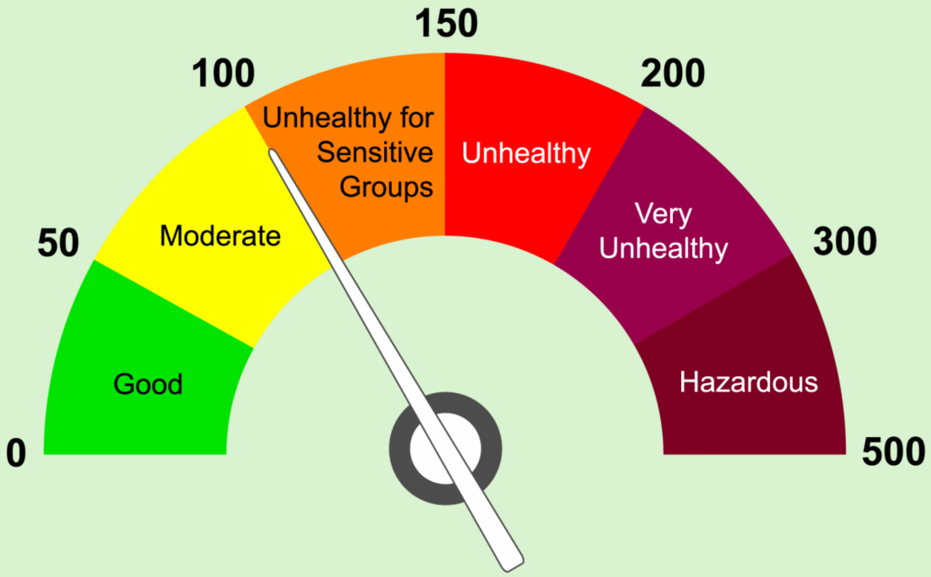
RESEARCH METHODS

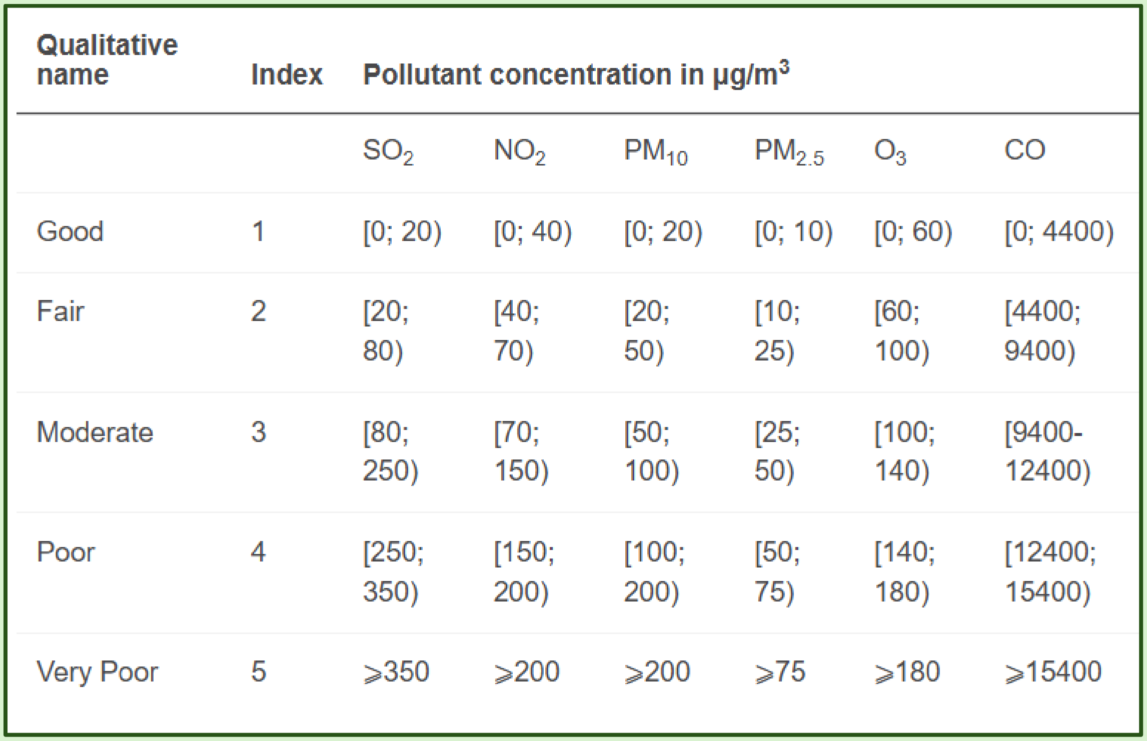
- **Data Collection and Preprocessing:** Initially, air quality data was gathered from Varna's two Air Monitoring Stations, "Chaika" and "Angel Kanchev", using official datasets and APIs like OpenAQ. There was quite a lot of missing or inconsistent values, which caused some issues. However, during the project, the data source was switched to an API provided by OpenWeatherMap. This decision was made due to OpenWeatherMap's comprehensive coverage and more recent data.
- **Exploratory Data Analysis:** Use statistical and visualization techniques (e.g., correlation analysis, line charts, bar charts) to understand pollutant behaviours and their relationships with AQI. Identify key pollutants that have the strongest correlation with air quality levels.
- **Regression Modelling and Evaluation:** Develop predictive models using algorithms such as Linear Regression, Random Forest, etc. Train and test models on the cleaned dataset, then evaluate performance with metrics like MSE, RMSE and R^2 to ensure accurate AQI predictions.

ANALYTIC APPROACH

TARGET VARIABLE

The target variable is the **Air Quality Index (AQI)**, which is calculated by first converting pollutant concentrations into sub-indices using standardized breakpoint tables.





Qualitative name	Index	Pollutant concentration in $\mu g/m^3$					
		SO_2	NO_2	PM_{10}	$PM_{2.5}$	O_3	CO
Good	1	[0; 20)	[0; 40)	[0; 20)	[0; 10)	[0; 60)	[0; 4400)
Fair	2	[20; 80)	[40; 70)	[20; 50)	[10; 25)	[60; 100)	[4400; 9400)
Moderate	3	[80; 250)	[70; 150)	[50; 100)	[25; 50)	[100; 140)	[9400-12400)
Poor	4	[250; 350)	[150; 200)	[100; 200)	[50; 75)	[140; 180)	[12400; 15400)
Very Poor	5	⩾350	⩾200	⩾200	⩾75	⩾180	⩾15400

Parameters that do not affect the AQI: NH_3 and NO

TYPE OF PROBLEM

Regression – the goal is to predict a continuous AQI value from environmental data.

POTENTIAL MACHINE LEARNING MODELS

- **Linear Regression:** Used as a baseline model to capture linear relationships between pollutants and AQI.
- **Ridge Regression:** Like Linear Regression but with regularization to prevent overfitting.
- **Lasso Regression:** Introduced L1 regularization to encourage sparsity in the feature set.
- **Random Forest Regressor:** A robust ensemble method that models complex, non-linear relationships and handles interactions between pollutants.
- **Extra Trees Regressor:** A variant of Random Forest that splits nodes randomly.
- **Bagging Regressor:** Combines multiple weak learners to improve stability and accuracy.
- **K-Nearest Neighbors (KNN):** A distance-based model that relies on local patterns in the data.
- **Support Vector Regressor (SVR):** Effective for smaller datasets.
- **Gradient Boosting Regressor:** Combines weak learners iteratively to optimize performance.
- **XGBoost:** A gradient boosting algorithm optimized for speed and performance.
- **LightGBM (LGBM):** Known for efficiency and accuracy.
- **CatBoost:** Particularly effective with categorical data and yielded competitive results.

DEFINING SUCCESS

Success will be defined by:

- Low prediction errors measured by Mean Squared Error (MSE) and Root Mean Squared Error (RMSE).
- High coefficient of determination (R^2 score) indicating strong explanatory power.
- Visual alignment of predicted vs. actual AQI values through plots.
- Model robustness.

DATA REQUIREMENTS

DEFINE OBJECTIVES

- Predict AQI for Varna using pollutant concentration data.
- Understand the influence of each pollutant on air quality.
- Provide actionable insights to stakeholders for urban air quality management.

DATA CHARACTERISTICS

- Pollutant concentration values for CO, NO, NO_2, O_3, SO_2, PM2_5, PM10 and NH_3 (in µg/m^3).
- Timestamps with both UTC and local time.

DATA SOURCES

- Primary Data Source: Air quality data was collected using the **[*“OpenWeatherMap Air Pollution API”*]**, which provides comprehensive environmental metrics including key pollutants and air quality indices.
- Supplementary Information: The API was selected to ensure continuous and reliable access to up-to-date air quality data for Varna, replacing earlier data collection efforts from local monitoring stations like “Chaika” and “Angel Kanchev”.

DATA LEGALITY AND ETHICS

- Ensure compliance with Bulgarian data privacy laws and OpenWeatherMap usage policies.
- Use data responsibly to inform public health without causing undue alarm.
- Provide transparency about data sources and model limitations.

DATA DIVERSITY

- Include multiple pollutants to cover spatial and temporal variability.
- Capture data across various times of day to improve generalizability.

VERSION CONTROL

Using GitHub for code and dataset versioning: **["*VarnaAir*"]**

ITERATIVE PROCESS

It is important to evaluate model performance, refine preprocessing steps, update the dataset and continuously incorporate feedback.

CONCLUSION

This project, **VarnaAir**, is an opportunity to apply data science and machine learning techniques to a real-world environmental challenge that affects my hometown deeply. By predicting the Air Quality Index using local environmental data, I hope to provide valuable insights for citizens and policymakers to improve urban health and quality of life. The project will also help me grow my technical skills in regression analysis and data handling, fulfilling important learning outcomes while contributing positively to the community I care about.