VARNA AIR

PROJECT PROPOSAL

08.06.2025

IVET KALCHEVA

CONTENTS

Introduction 3

What 3

Why 3

Who 3

When 3

How 3

Domain Understanding 4

Research Question 4

Exploratory Research 4

Research Methods 4

Analytic Approach 5

Target Variable 5

Type of Problem 5

Potential Machine Learning Models 5

Defining success 5

Data Requirements 6

Define Objectives 6

Data Characteristics 6

Data Sources 6

Data Legality and Ethics 6

Data Diversity 6

Version Control 6

Iterative Process 7

Conclusion 7

INTRODUCTION

WHAT

This project aims to predict the Air Quality Index (AQI) for Varna, Bulgaria, using regression models and data from two Air Monitoring Stations (AMS) – "*Chaika*" and "*Angel Kanchev*". The primary focus is on leveraging environmental data such as pollutant levels (CO, NO, NO_2, O_3, PM10, PM2.5 and SO_2) 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. This project also allows me to develop a deeper understanding of data handling and analysis, aligning with my learning outcome goals.

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, PM10, PM2.5, SO_2) have the strongest correlation with AQI values?
- How do pollutant levels vary spatially between the "Chaika" and "Angel Kanchev" monitoring stations?
- Are there any temporal trends or seasonal variations in pollutant concentrations?

RESEARCH METHODS

- **Data Collection and Preprocessing:** Gather air quality data from Varna's two Air Monitoring Stations ("Chaika" and "Angel Kanchev") via official datasets or APIs like OpenAQ. Clean the data by handling missing or inconsistent values, standardizing units and converting timestamps to local time for consistency.
- **Exploratory Data Analysis:** Use statistical and visualization techniques (e.g., correlation analysis, line charts, boxplots) 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. The AQI is then the maximum of these sub-indices:

$$\boldsymbol{AQI = max(I_{co}, I_{NO2}, I_{O3}, I_{PM2.5}, I_{PM10}, I_{SO2})}$$

where each $I_{pollutant}$ is the sub-index corresponding to the pollutant concentration.

TYPE OF PROBLEM

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

POTENTIAL MACHINE LEARNING MODELS

- **Linear Regression:** Baseline model to capture linear relationships between pollutants and AQI.
- **Random Forest Regressor:** Can model complex, non-linear relationships and handle interactions between pollutants.

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 tested on data from different times and locations.

DATA REQUIREMENTS

DEFINE OBJECTIVES

- Predict AQI for Varna using pollutant concentration data from monitoring stations.
- 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, PM10, PM2.5, SO_2 (in µg/m^3).
- Timestamps with both UTC and local time.
- Location data specifying which AMS station collected the data.
- Metadata: station type, ownership, whether data is from a mobile or fixed monitor.

DATA SOURCES

- Data collected from Varna's Air Monitoring Stations: "Chaika" and "Angel Kanchev".
- Supplementary data from OpenAQ and Bulgarian environmental agencies (if available).

DATA LEGALITY AND ETHICS

- Ensure compliance with Bulgarian data privacy laws and OpenAQ 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 and both AMS stations to cover spatial and temporal variability.
- Capture data across various times of day and weather conditions 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.