**Project Title:** Predicting Air Quality for Sustainable Urban Living

**Team Members**

1. Dawit Teklebrehan
2. Shimels Tilahun
3. Robel Ermiyas
4. Robel Roba
5. Tewodros Gebretsadk
6. Negassa Retta

**Date of Submisson: April 18, 2025**

Table of Contents

[**Concept Note** 1](#_Toc197748708)

[1. Project Overview 1](#_Toc197748709)

[2. Objectives 1](#_Toc197748710)

[3. Background 1](#_Toc197748711)

[4. Methodology 2](#_Toc197748712)

[5. Architecture Design Diagram 2](#_Toc197748713)

[6. Data Sources 3](#_Toc197748714)

[7. Literature Review 3](#_Toc197748715)

[**Implementation Plan** 4](#_Toc197748716)

[1 Technology Stack 4](#_Toc197748717)

[2. Timeline 4](#_Toc197748718)

[**Task Distribution Matrix** 5](#_Toc197748719)

[3. Milestones 5](#_Toc197748720)

[4. Challenges and Mitigations 6](#_Toc197748721)

[**Challenge 1: Data Quality and Availability** 6](#_Toc197748722)

[**Challenge 2: Model Performance and Accuracy** 6](#_Toc197748723)

[**Challenge 3: Technical Constraints and Resource Limitations** 6](#_Toc197748724)

[**Challenge 4: Limited User Access to Technology** 6](#_Toc197748725)

[5. Ethical Considerations 7](#_Toc197748726)

[6. References 8](#_Toc197748727)

# ****Concept Note****

## 1. Project Overview

This capstone project, titled *"Predicting Air Quality for Sustainable Urban Living,"* aligns with multiple **United Nations Sustainable Development Goals (SDGs)**, particularly **SDG 3 (Good Health and Well-being)**, **SDG 11 (Sustainable Cities and Communities)**, and **SDG 13 (Climate Action)**. The project addresses the escalating problem of urban air pollution, which significantly impacts public health and environmental sustainability. In rapidly urbanizing regions like Addis Ababa, air pollution stems from vehicular emissions, industrial activities, and biomass burning. The proposed solution leverages machine learning to provide real-time and predictive insights into air quality levels. By forecasting air pollution trends, the project seeks to inform urban planning, enhance public health outcomes, and support policy interventions aimed at reducing exposure to harmful pollutants.

## 2. Objectives

* **Develop a predictive machine learning model** capable of estimating future air quality indices (AQIs) based on historical and real-time environmental data.
* **Analyze historical pollution trends** to identify temporal and spatial patterns in urban air quality.
* **Design an accessible visualization dashboard** that communicates air quality forecasts to the general public and policymakers in an intuitive format.
* **Facilitate data-driven decision-making** in urban governance and health sectors to implement timely interventions in pollution-prone zones.
* **Contribute to academic and technical knowledge** in the field of environmental monitoring using artificial intelligence.

## 3. Background

Urban air pollution poses one of the most serious environmental health risks worldwide, contributing to approximately 7 million premature deaths annually, according to the World Health Organization (WHO). In Ethiopia, and particularly Addis Ababa, urban expansion, increased motorization, and unregulated industrial emissions have exacerbated air quality concerns. Traditional air quality monitoring approaches are limited in both spatial coverage and responsiveness. While some initiatives have deployed sensors and mobile-based monitoring, they often lack predictive capabilities. A machine learning approach offers a scalable, cost-effective, and adaptive framework to forecast pollution levels using diverse environmental data sources. Predictive analytics enable timely alerts and interventions, providing city officials and citizens with foresight into pollution trends and potential health risks.

## 4. Methodology

The project will employ **supervised machine learning techniques** with a focus on **regression-based models** to predict air pollutant concentrations (e.g., PM2.5, PM10, NO2). Models under consideration include:

* **Random Forest Regressor:** Due to its robustness against overfitting and ability to handle nonlinear relationships.
* **Gradient Boosting Machines (GBM):** For its accuracy in capturing complex feature interactions.
* **Support Vector Regression (SVR):** Particularly useful for small- to medium-sized datasets with high-dimensional features.

Data preprocessing will involve handling missing values, encoding time-based features (e.g., hour of day, day of week), normalizing sensor data, and aggregating meteorological parameters. Evaluation metrics will include **Mean Absolute Error (MAE)**, **Root Mean Square Error (RMSE)**, and **R-squared (R²)**. The development stack will leverage **Python, Scikit-learn, Pandas, NumPy, Matplotlib**, and **Streamlit** for model visualization and interaction.

## 5. Architecture Design Diagram

A high-level architecture of the project includes the following components:

1. **Data Collection Layer:**
   * Sources include OpenAQ API, World Air Quality Index, NASA Earth observations.
   * Ingests real-time and historical data on pollutants and meteorological conditions.
2. **Data Preprocessing Layer:**
   * Cleans raw data, handles missing values, scales features, and performs feature engineering.
3. **Model Training Layer:**
   * Applies regression algorithms to train predictive models using labeled data.
4. **Model Evaluation Layer:**
   * Validates model performance with historical data using error metrics.
5. **Prediction Layer:**
   * Generates future AQI values for short-term forecasts.
6. **Visualization Layer:**
   * Provides interactive web dashboard for end-users via Streamlit, including color-coded maps, charts, and real-time forecasts.

Each layer is modular and designed to ensure scalability, reusability, and easy integration into smart city platforms.

## 6. Data Sources

The data used in this project will be sourced from open, trusted platforms including the **OpenAQ platform**, **World Air Quality Index (WAQI)**, and **NASA Earth Observations**. These sources provide granular, real-time data on pollutants such as PM2.5, PM10, NO2, SO2, CO, and O3, as well as meteorological variables like humidity, temperature, and wind speed. Preprocessing steps include temporal alignment, normalization, outlier removal, and feature extraction to enrich the training dataset for modeling purposes.

## 7. Literature Review

Previous studies have demonstrated the efficacy of machine learning models such as Random Forest and LSTM in accurately predicting air pollution levels. Research emphasizes the integration of temporal and meteorological variables for more robust performance. Additionally, data visualization tools have proven critical in enhancing public engagement and awareness. This project extends existing work by combining multiple regression models, real-time prediction capabilities, and an interactive dashboard to support both public health initiatives and urban planning strategies.

# Implementation Plan

## 1 Technology Stack

The implementation of the project will rely on the following technology stack:

* **Programming Language:** Python
* **Libraries and Frameworks:**
  + **Scikit-learn** – for machine learning model development
  + **Pandas and NumPy** – for data manipulation and numerical computation
  + **Matplotlib and Seaborn** – for data visualization
  + **TensorFlow/Keras (optional)** – for deep learning experiments
  + **Streamlit** – for building interactive dashboards
* **Data Sources/APIs:**
  + OpenAQ API
  + World Air Quality Index API
  + NASA Earth Observation datasets
* **Development Tools:**
  + Google Colab / Jupyter Notebook
  + GitHub – for version control
* **Deployment Tools:**
  + Streamlit Cloud / Heroku – for web app deployment

## 2. Timeline

|  |  |  |  |
| --- | --- | --- | --- |
| **Phase** | **Task** | **Duration** | **Deadline** |
| Phase 1 – Planning | Requirement analysis and literature review | 1 week | Week 1 |
| Phase 2 – Data Collection | Acquire, clean, and preprocess datasets | 1 week | Week 2 |
| Phase 3 – Modeling | Feature engineering and model training | 2 weeks | Week 3–4 |
| Phase 4 – Evaluation | Hyperparameter tuning and testing | 1 week | Week 5 |
| Phase 5 – Deployment | Build and deploy Streamlit dashboard | 1 week | Week 6 |
| Phase 6 – Documentation | Write final report and prepare presentation | 1 week | Week 7 |

**Task Distribution Matrix**

|  |  |
| --- | --- |
| **Team Member** | **Responsibility** |
| Dawit Teklebrehan | Model selection, algorithm tuning |
| Shimels Tilahun | Data preprocessing, visualization |
| Robel Ermiyas | Dataset exploration, documentation |
| Robel Roba | Frontend development (Streamlit), API integration |
| Tewodros Gebretsadk | Literature review, evaluation metrics |
| Negassa Retta | Deployment, testing, final report |

## 3. Milestones

The following milestones have been identified to mark significant progress points in the project:

* **Milestone 1: Completion of Requirement Analysis and Literature Review**  
  *Expected Date:* End of Week 1  
  Deliverables: Finalized project scope, annotated bibliography, literature summary.
* **Milestone 2: Data Acquisition and Preprocessing Completed**  
  *Expected Date:* End of Week 2  
  Deliverables: Cleaned dataset, feature matrix, handling of missing values.
* **Milestone 3: Completion of Initial Model Training**  
  *Expected Date:* End of Week 4  
  Deliverables: Trained base models (Random Forest, SVR, GBM), preliminary results.
* **Milestone 4: Model Optimization and Evaluation Finalized**  
  *Expected Date:* End of Week 5  
  Deliverables: Optimized model with selected hyperparameters, evaluation metrics (MAE, RMSE, R²).
* **Milestone 5: Web Application Deployment**  
  *Expected Date:* End of Week 6  
  Deliverables: Fully functioning Streamlit dashboard hosted online.
* **Milestone 6: Final Report and Presentation Submitted**  
  *Expected Date:* End of Week 7  
  Deliverables: PDF Report, Slide Deck, and Presentation Recording.

## 4. Challenges and Mitigations

**Challenge 1: Data Quality and Availability**

* *Issue:* Data may have missing or inconsistent values, or may not be available for all pollutants.
* *Mitigation:* Use multiple reliable data sources (OpenAQ, WAQI, NASA), apply robust preprocessing (imputation, filtering), and generate synthetic samples using data augmentation if needed.

**Challenge 2: Model Performance and Accuracy**

* *Issue:* Regression models may underperform if input features are not well-structured.
* *Mitigation:* Apply advanced feature engineering, ensemble modeling, and hyperparameter tuning. Perform cross-validation and use error metrics to guide improvements.

**Challenge 3: Technical Constraints and Resource Limitations**

* *Issue:* Limited compute resources may slow training or limit model complexity.
* *Mitigation:* Utilize Google Colab with GPU acceleration. Limit model complexity or batch sizes. Deploy lightweight models for real-time use.

**Challenge 4: Limited User Access to Technology**

* *Issue:* Target users in rural or low-tech environments may have limited internet or smartphone access.
* *Mitigation:* Ensure the app is mobile-friendly, explore SMS-based alerts in future phases, and support multilingual interfaces.

## 5. Ethical Considerations

* **Data Privacy:** The project will not collect personal data from users. All datasets will be open-source and anonymized. Data used will be stored securely with restricted access.
* **Bias and Fairness:** The model may perform differently across regions due to variations in environmental data. We will test and report performance across different geographic areas and seasons to ensure fairness.
* **Transparency:** Model performance, training process, and limitations will be clearly communicated to stakeholders through documentation and dashboard tooltips.
* **Community Impact:** The project is designed to empower users and improve public health. All visualizations and outputs will be presented in a way that promotes understanding rather than fear or misinformation.

## 6. References

1. OpenAQ API Documentation – https://docs.openaq.org/
2. World Air Quality Index – https://waqi.info/
3. NASA Earth Observations – https://neo.sci.gsfc.nasa.gov/
4. WHO – Ambient air pollution: Health impacts – https://www.who.int/
5. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
6. Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society*, Series B.
7. Zhang, Y., & Ding, Y. (2017). A deep learning approach to air quality prediction based on meteorological data. *Sensors*, 17(6), 1348.
8. Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.