**📄 Project Proposal: Air Quality Index Prediction and Classification**

**1. Project Title:**

**Prediction and Classification of Air Quality Index (AQI) Using Machine Learning Techniques**

**2. Background and Problem Statement:**

Air pollution poses a serious threat to environmental sustainability and human health. Monitoring and forecasting air quality is essential to reduce health risks and to inform the public about necessary precautions.  
Traditional AQI monitoring systems often rely on expensive, large-scale sensor networks and complex environmental models.  
Moreover, standard AQI formulas depend heavily on the availability of complete pollutant data such as PM2.5, PM10, O₃, CO, NO₂, and SO₂.

However, due to data limitations, especially in certain regions, not all pollutant concentrations are available. This project addresses this gap by creating a **simplified, predictive AQI model** using available sensor data (CO, NO₂, O₃) and developing a reliable machine learning system that:

* Predicts a numerical AQI value.
* Classifies the air quality into meaningful categories such as Good, Moderate, Unhealthy, etc.

This solution aims to be **lightweight, accessible**, and **practical** for areas with limited data resources.

**3. Objectives:**

* Develop a simplified AQI calculation formula based on available pollutant data.
* Build a Regression model to predict AQI values from pollutant concentrations.
* Build a Classification model to categorize air quality based on AQI ranges.
* Evaluate and validate the models using statistical metrics.
* Deploy a simple and interactive web application for real-time AQI prediction and classification.

**4. Scope of Work:**

| **Phase** | **Deliverables** |
| --- | --- |
| Data Preprocessing | Clean and prepare the dataset for modeling |
| Feature Engineering | Design a Simple AQI formula based on weighted pollutant data |
| Exploratory Data Analysis (EDA) | Visualize and understand data distributions and correlations |
| Model Building | Train Regression and Classification models |
| Evaluation | Analyze model performance using error metrics and confusion matrix |

**5. Dataset Description:**

| **Attribute** | **Description** |
| --- | --- |
| Date | Date of observation |
| Time | Time of observation |
| CO(GT) | Carbon Monoxide concentration (mg/m³) |
| PT08.S1(CO) | CO sensor output |
| NMHC(GT) | Non-Methane Hydrocarbons concentration (µg/m³) |
| C6H6(GT) | Benzene concentration (µg/m³) |
| PT08.S2(NMHC) | NMHC sensor output |
| NOx(GT) | Nitrogen Oxides concentration (ppb) |
| PT08.S3(NOx) | NOx sensor output |
| NO2(GT) | Nitrogen Dioxide concentration (µg/m³) |
| PT08.S4(NO2) | NO2 sensor output |
| PT08.S5(O3) | Ozone sensor output |
| T | Temperature (°C) |
| RH | Relative Humidity (%) |
| AH | Absolute Humidity (g/m³) |

**Source:** Air Quality Dataset (Multisite Monitoring) — Public Repository.

**6. Methodology:**

**Data Cleaning and Preprocessing:**

* Load the dataset with proper separator (; instead of ,).
* Replace decimal commas with decimal points.
* Convert columns to appropriate data types (floats).
* Remove missing values and drop unnecessary columns.

**Feature Engineering:**

* Define a custom AQI based on:

Simple AQI=(0.3×CO(GT))+(0.4×NO2(GT))+(0.3×PT08.S5(O3))\text{Simple AQI} = (0.3 \times \text{CO(GT)}) + (0.4 \times \text{NO2(GT)}) + (0.3 \times \text{PT08.S5(O3)})

* Add an AQI\_Category label based on AQI thresholds.

**Exploratory Data Analysis (EDA):**

* Correlation heatmaps between features and target.
* Histograms to check AQI distribution.
* Boxplots if needed for outlier detection.

**Modeling:**

| **Task** | **Approach** |
| --- | --- |
| AQI Prediction | Linear Regression Model |
| AQI Categorization | Random Forest Classifier |

* Data split: 80% training / 20% testing.
* Hyperparameter tuning if necessary.

**Evaluation Metrics:**

| **Model** | **Metrics** |
| --- | --- |
| Regression | Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), R² Score |
| Classification | Accuracy, Confusion Matrix, Precision, Recall, F1 Score |

**Deployment:**

* Create a user-friendly **Streamlit Web App**.
* Input: CO, NO₂, O₃ values.
* Output: Predicted AQI value + AQI Category.

**7. Tools and Technologies:**

* **Python 3.x**
* **GOOGLE COLLAB** for model development.
* **Pandas, NumPy** for data manipulation.
* **Matplotlib, Seaborn** for visualization.
* **scikit-learn** for machine learning modeling.

**8. Expected Outcomes:**

* A predictive model that can forecast AQI with reasonable accuracy.
* A classification model that informs users whether the air quality is good or hazardous.

**9. Timeline (Example for 4 weeks):**

| **Week** | **Activities** |
| --- | --- |
| 1 | Data cleaning, preprocessing, feature engineering |
| 2 | EDA, Model building (Regression + Classification) |
| 3 | Model evaluation, optimization, Streamlit interface |

**10. Conclusion:**

This project represents a meaningful attempt to build a cost-effective, intelligent AQI prediction and classification tool in situations where complete environmental data may not be available.  
The combination of data science techniques and lightweight web deployment ensures that the project has **both technical depth and practical applicability**.

The proposed model can serve as a **starting point for more complex environmental monitoring systems** in the future.