**Project Proposal**

**Title:** “Analysis and Early Detection of Respiratory Illness-Related Cases Using UK Air Quality and Hospital Data”

**Region Focus:** West Midlands

**Data Period:** June 2024 – June 2025

**1. Objective**

This project aims to analyze patterns and develop early warning models for surges in specific respiratory illnesses — acute bronchiolitis, influenza-like illness (ILI), scarlet fever, and acute respiratory infections (ARI) — in the West Midlands region. By linking environmental air quality data with hospital case trends, this system seeks to support timely public health responses and improve resource readiness in healthcare systems.

**2. Data Sources**

**Health / Respiratory Illness Data**

Source: UK Health Security Agency (UKHSA) Dashboard

Illness Types Tracked:

* Acute Bronchiolitis
* Influenza-like Illness (ILI)
* Scarlet Fever
* Acute Respiratory Infections (ARI)

Use: Daily counts of diagnosed cases in West Midlands

**Air Quality Data**

Source: DEFRA UK-AIR Database

Pollutants Tracked:

* PM2.5
* PM10
* Black Carbon
* Blue Particulate Matter
* Yellow Particulate Matter
* Red Particulate Matter
* Green Particulate Matter
* Nitric Oxide (NO)
* Nitric Dioxide (NO₂)
* Ozone (O₃)

Use: Daily pollutant levels matched with illness trends, Feature engineering with lags (1-day, 3-day, 7-day) and moving averages.

**3. Timeline**

**Week 1–2: Data Acquisition & Preprocessing**

* Download and clean illness case counts and pollutant data.
* Align datasets by date and geography.
* Handle missing values and normalize formats.
* Create engineered features: lagged pollutants, exposure windows, temperature/humidity (optional).

**Week 3–4: Exploratory Data Analysis (EDA)**

* Visualize time-series trends in illnesses and pollutants.
* Correlate pollution levels with illness spikes.
* Identify seasonality and spatial variation.
* Compare severity across illness categories.

**Week 5: Modeling, Interpretation & Validation**

* Model formulation: Time-based regression for case prediction.
* Algorithms: Random Forest, XGBoost, SARIMAX, Multi-output Regression.
* Split data temporally for training/testing.
* Evaluate with RMSE, MAE, and R².
* Use SHAP and feature importance to interpret pollutant impact.
* Validate across seasonal windows for consistency.

**Week 6: Visualization & Reporting**

* Develop visual outputs: weekly trends, pollution vs. illness overlays, actual vs. predicted surges.
* Generate regional risk indicators.
* Compile a concise final report:
  + Executive Summary.
  + Illness-specific Results.
  + Public Health Recommendations.
* Publish GitHub repository with code, documentation, and results.

**4. Tools & Technologies**

* Python: pandas, scikit-learn, XGBoost, statsmodels, seaborn, matplotlib.
* SQL: for querying tabular hospital records.
* Power BI / Tableau: for data storytelling.

**5. Real-World Applications**

**Healthcare Planning**

* Challenge: Sudden spikes in illness cases strain hospital capacity
* Solution: Early warning predictions for each condition to:
  + 1. Pre-stock paediatric and respiratory treatments.
    2. Pre-warn A&E teams on likely spikes in patient visits.

**Public Health Alerts**

* Predict rising risks for conditions like ILI or ARI during pollution peaks.
* Enable local authorities to alert schools, care homes, and the public consciousness.
* Suggest preventive measures (e.g., mask use, limit outdoor activity).

**Environmental Policy Impact**

* Strengthen the case for regulating pollutants like NO₂, PM2.5, and Black Carbon.
* Support initiatives for clean air zones and emissions reduction in the West Midlands.