Explainable AI - Assignment 2 Report

This report presents the analysis of air quality data (PRSA Data, 2010–2014) using a Random Forest Regressor model and Explainable AI techniques (SHAP). The primary objective is to predict PM2.5 concentrations and interpret the influence of different meteorological factors on air pollution levels.

Dataset:
- Source: PRSA Data (Beijing, 2010–2014)
- Records: Hourly observations
- Features: PM2.5, TEMP, DEWP, PRES, Wind (cbwd, Iws), Snow (Is), Rain (Ir)
- Preprocessing: Dropped time columns, handled missing values, one-hot encoded wind direction, sampled 2000 records.

Methodology:
- Model: Random Forest Regressor (n_estimators=50)
- Train-Test Split: 80:20
- Explainability: SHAP (summary, force, and waterfall plots) applied on test samples to understand feature contributions.

Results:

Metric	Value
RMSE	73.07
MSE	5339.28
MAE	51.68
R^2	0.36

Top 5 Influential Features:
1. Temperature (TEMP)
2. Dew Point (DEWP)
3. Wind Speed (Iws)
4. Pressure (PRES)
5. Wind Direction SE (cbwd_SE)

Discussion:
- Weather conditions such as temperature, dew point, wind speed, and pressure strongly influence PM2.5 levels.
- Rain and snow events also help in reducing pollution levels.
- The model performance ($R^2 = 0.36$) suggests underfitting; more features or advanced models (XGBoost, LSTM) may improve prediction accuracy.

Conclusion:
Random Forest with SHAP provides interpretability in PM2.5 prediction, highlighting key meteorological drivers of air quality. Future improvements could include using deep learning models and incorporating additional environmental variables for better performance.