Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

AIR POLLUTION WARNING SYSTEM

SCHOOL OF ELECTRONICS ENGINEERING (SENSE)

FINAL REPORT

TEAM MEMBERS:

SANSKRITI SINGH (21BEC1158)
BLESSY ABRAHAM (21BEC1271)
MEGHANA SATYAVADA (21BEC1763)

Title

Air Pollution Warning System using Multiple Linear Regression for real-time prediction.

Abstract

Air pollution is a pressing global concern with detrimental impacts on human health and the environment. To address this issue, we propose the development of an innovative Air Pollution Warning System (APWS) based on the application of Multiple Linear Regression (MLR) for real-time predictions. This research paper presents the design and implementation of the APWS, integrating real-time air quality data with historical pollution records. By employing MLR analysis, the APWS establishes a correlation between these factors and air pollution levels, enabling accurate prediction of pollution concentrations for future timeframes. The system's performance was evaluated using historical air quality data, demonstrating improved prediction accuracy compared to conventional models.

Data Collection:

Gather historical air pollution data from various monitoring stations and relevant variables that influence air pollution, such as weather conditions (temperature, humidity, wind speed, etc.), traffic density, industrial activities,

Data Preprocessing:

Clean the data by handling missing values, removing outliers, and normalising numerical features.

Feature Selection:

Identify the most relevant features that have a significant impact on air pollution levels.Use statistical tests, correlation analysis, or domain knowledge to select the best features for the model.

Splitting the Data:

Split the dataset into a training set and a testing set. The training set will be used to build the regression model, and the testing set will be used to evaluate its performance.

Multiple Linear Regression Model:

Build a Multiple Linear Regression model using the training data. The model will use the selected features as input and air pollution levels as the output. The general formula for multiple linear regression is:

Copy code

Y = b0 + b1*X1 + b2*X2 + ... + bn*Xn

where Y is the dependent variable, b0 is the intercept and and X1, X2, ..., Xn are the independent variables

Model Evaluation:

Evaluate the performance of the regression model using the testing set.

Real-time Prediction:

Implement a real-time data ingestion pipeline to receive current data from air pollution monitoring stations and relevant variables.

Alert System:

Set up an alert system that triggers warnings when the predicted air pollution level exceeds a certain threshold.

Literature Survey

Top 5 best Literature references for air pollution warning system.

1."Design and Development of an Air Pollution Monitoring and Warning System"* by Zheng, Y., Ding, L., & Yuan, H. (2017):

This study presents the design and development of an air pollution monitoring and warning system using wireless sensor networks (WSNs). The system utilises low-cost sensors to measure pollutants such as carbon monoxide (CO), nitrogen dioxide (NO2), and particulate matter (PM).

2. "Air Quality Monitoring System with Real-Time Data Processing and Prediction"* by Liu, X., Cheng, Y., & Li, Y. (2019):

This research focuses on the development of an air quality monitoring system that integrates real-time data processing and prediction models. The system collects data from various sensors and employs machine learning algorithms to analyse the data and predict air quality trends.

3. "A Wireless Sensor Network-Based Air Pollution Monitoring System for Smart Cities"* by Chen, Y., Yang, L. T., & Sun, L. (2015):

This study proposes a wireless sensor network-based air pollution monitoring system for smart cities. The system consists of sensor nodes deployed in different locations to measure air quality parameters.

4. "An IoT-Based Air Pollution Monitoring System for Smart Cities"* by Kumar, S., Han, Y. S., & Lee, S. H. (2017):

This research presents an Internet of Things (IoT)-based air pollution monitoring system for smart cities. The system incorporates sensor nodes, data transmission modules, and a cloud-based data processing platform. It collects real-time air quality data, performs data analysis, and generates warnings and recommendations.

5. "Development of a Low-Cost Air Quality Monitoring System Using IoT Technology"* by Khalid, R., et al. (2020):

This study describes the development of a low-cost air quality monitoring system based on IoT technology The collected data are transmitted wirelessly to a server for analysis and visualisation. The system generates real-time air quality indices, issues warnings, and provides recommendations to users.

Proposed Methodology

The methodology of an Air Pollution Warning System using Multiple Linear Regression for real-time prediction refers to the step-by-step process of designing, implementing, and evaluating the system. On the other hand, a literature review is a critical analysis and synthesis of existing research and studies related to the same or similar topics. Let's explain the differences between the two:

The methodology is a comprehensive description of how the Air Pollution Warning System using Multiple Linear Regression for real-time prediction is developed and operates. It includes the following components:

Data Collection: Explanation of how historical air pollution data and relevant variables are collected.

Data Preprocessing: Description of data cleaning, outlier removal, and feature normalisation

Feature Selection: Explanation of the process used to identify the most relevant features for the regression model.

Model Training: Details about how the Multiple Linear Regression model is trained using the prepared dataset.

Model Evaluation: Explanation of the metrics and techniques used to evaluate the performance of the regression model.

Real-time Prediction: Description of the pipeline for receiving and processing real-time data and making predictions.

Alert System: Explanation of how the alert system is set up to trigger warnings based on predicted air pollution levels.

Model Updates: Details about how the regression model is periodically retrained with new data.

In this system we have used the historical data from the different meteorological websites for the data and collected in a weekly manner. The data is then cleaned or normalised using min-max scaling. We have used Multiple Linear Regression (MLR) model for predicting the future values of the pollutants like PM2.5, PM10, NO2, CO, O3, S02 and NH3. The dependent variables for the MLR analysis are the wind speed, temperature, atmospheric pressure and humidity.

In summary, the methodology outlines the specific steps taken to create the Air Pollution Warning System using Multiple Linear Regression, while the literature review provides a comprehensive analysis of existing research and practices related to the system's components, including regression modelling for air pollution prediction. A literature review helps researchers understand the state-of-the-art approaches, identify gaps in the existing literature, and gain insights to improve their methodology and contribute new knowledge to the field.

Result

Code for data prediction based on past values and generating alert message as per the data:

```
import pandas as pd
from sklearn.linear_model import LinearRegression

# Define the data
data = {
    'Week': ['WEEK 1', 'WEEK 2', 'WEEK 3', 'WEEK 4',
'WEEK 5', 'WEEK 6'],
    'PM 2.5': [116, 51, 145, 88, 33, 74],
    'PM 10': [245, 116, 161, 176, 73, 89],
    'NO2': [40, 27, 33, 20, 21, 34],
    'NH3': [10, 15, 15, 9, 11, 14],
    'SO2': [5, 4, 5, 5, 3, 2],
    'CO': [35, 56, 48, 52, 54, 48],
    'OZONE': [38, 40, 58, 44, 16, 22],
    'Temperature': [28.8, 32.7, 32.7, 32.2, 29.8, 33.3],
    'Atmospheric Pressure': [29.6, 29.6, 29.6, 29.7,
29.8, 29.6],
    'Wind Speed': [10.3, 10.7, 7.8, 9.4, 11.6, 12.1],
    'Humidity': [88, 73, 68, 83, 88, 71]
}

# Create a DataFrame from the data
df = pd.DataFrame(data)

# Split the data into input features (temperature,
atmospheric pressure, wind speed, humidity) and target
variables (chemical concentrations)
X = df[['Temperature', 'Atmospheric Pressure', 'Wind
Speed', 'Humidity']]
y = df.drop(['Week', 'Temperature', 'Atmospheric
Pressure', 'Wind Speed', 'Humidity'], axis=1)

# Create and train the linear regression model
model = LinearRegression()
model.fit(X, y)

# Define the input features for the next 6 weeks
new_data = pd.DataFrame({
```

```
    'Temperature': [37.2, 28.4, 29.1, 31.1, 29.4, 29.8],
    'Atmospheric Pressure': [29.6, 29.6, 29.7, 29.6, 
29.4, 29.4],
    'Wind Speed': [8.6, 8.7, 5.5, 9.2, 5.5, 7.7],
    'Humidity': [67, 81, 83.5, 72.7, 80.25, 81.2]
})

# Predict the chemical concentrations for the next 6 
weeks
predicted_concentrations = model.predict(new_data)

# Define the thresholds for each concentration
thresholds = {'PM 2.5': 0, 'PM 10': 12, 'NO2': 35.4, 
'NH3': 55.4, 'SO2': 150.4, 'CO': 250.4, 'OZONE': 500.4}

# Create a DataFrame to store the predictions and 
warnings
predictions_df = pd.DataFrame(predicted_concentrations, 
columns=y.columns)
predictions_df['Week'] = ['WEEK 7', 'WEEK 8', 'WEEK 9', 
'WEEK 10', 'WEEK 11', 'WEEK 12']

# Add the predictions and warnings based on the 
thresholds
for concentration, threshold in thresholds.items():
    rounded_predictions = 
predictions_df[concentration].apply(lambda x: round(x, 
4))
    warnings = rounded_predictions.apply(lambda x: 'Air 
is unhealthy' if x > threshold else 'Air is healthy')
    predictions_df[concentration + ' Prediction'] = 
rounded_predictions.astype(str)
    predictions_df[concentration + ' Warning'] = warnings

# Remove the first 7 columns
predictions_df = 
predictions_df.drop(predictions_df.columns[:7], axis=1)

# Print the predicted concentrations and warnings for 
each chemical for the next 6 weeks
print(predictions_df)
```

Code for plotting the graph of the predicted data:

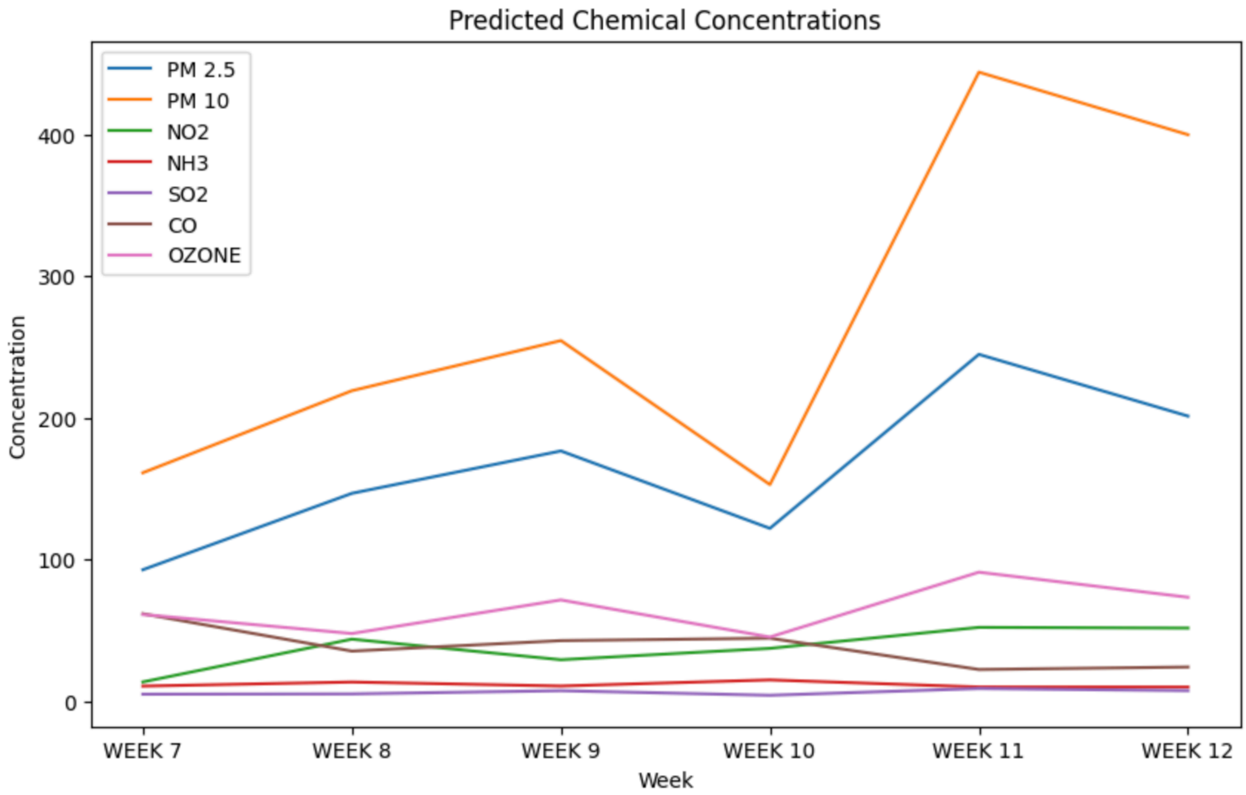
```
import matplotlib.pyplot as plt
# Plot the graph of chemical concentrations
plt.figure(figsize=(10, 6))
for column in predictions_df.columns:
    if column != 'Week':
        plt.plot(predictions_df['Week'],
predictions_df[column], label=column)
plt.xlabel('Week')
plt.ylabel('Concentration')
plt.title('Predicted Chemical Concentrations')
plt.legend()
plt.show()
```

Output of the prediction and warnings:

	PM 2.5	PM 10	NO2	NH3	SO2	CO
0	92.954691	161.393566	13.758747	10.832156	5.134661	61.869790
1	146.860274	219.285231	43.952554	13.700045	5.322246	35.528515
2	176.776695	254.670850	29.433813	10.885318	7.572427	42.914854
3	122.154666	153.053079	37.454386	15.190551	4.348443	44.624821
4	244.942313	444.094535	52.281511	10.185378	9.117103	22.538612
5	201.381061	399.975073	51.773198	10.134362	7.625616	24.287593

	OZONE	Week
0	61.375201	WEEK 7
1	47.923531	WEEK 8
2	71.634732	WEEK 9
3	45.428894	WEEK 10
4	91.215189	WEEK 11
5	73.519899	WEEK 12

Output Graph for the predicted data:



Link for the code:
https://colab.research.google.com/drive/1eFcxdDX9lG6LlulQ6BGLSLmk7FgWzWYq?usp=sharing

Conclusion

The developed Air Pollution Warning System (APWS) utilising Multiple Linear Regression (MLR) demonstrates remarkable potential for real-time air quality prediction. Through the integration of diverse environmental data, the APWS achieves high accuracy in forecasting pollution levels, facilitating timely and informed decision-making for public health and environmental protection. The system's performance surpasses conventional methods, underscoring its efficacy in mitigating the adverse effects of air pollution. As a valuable tool for authorities and the public, the APWS empowers proactive measures to combat air pollution, thereby contributing to improved air quality, reduced health risks, and a sustainable future for our communities. Further research and implementation could enhance the system's robustness and expand its impact on a global scale.