REAL TIME AIR QUALITY PREDICTION AND ANOMALY DETECTION

Dr.S.Ariffa Begum
CSE-Dept(Associate professor)
Kalasalingam Academy of Research and Education
Virudhunagar, Tamil Nadu, India

K.Naveen Kumar Reddy
Computer Science and Engineering
Kalasalingam Academy of Research and Education
Virudhunagar, Tamil Nadu, India
naveenkodumuru143@gmail.com

M. Rajesh Kumar Reddy
Computer Science and Engineering
Kalasalingam Academy of Research and Education
Virudhunagar, Tamil Nadu, India
marakarajeshkumarreddy@gmail.com

I. Siva Prasad Reddy
Computer Science and Engineering
Kalasalingam Academy of Research and Education
Virudhunagar, Tamil Nadu, India
xmirack@gmail.com

***Abstract*— When it comes to addressing the growing concerns about environmental health and safety, real-time air quality forecast and anomaly detection are essential. This study describes a robust system that effectively tracks, analyses, and forecasts air quality parameters in real-time using a combination of cutting-edge sensors, sophisticated data analytics, and machine learning algorithms. Utilising a wide range of environmental sensors, the system continuously monitors pollutants such as carbon monoxide, nitrogen dioxide, sulphur dioxide, particulate matter (PM2.5 and PM10), and ozone. Modern data analytics techniques are then used to process and analyse the gathered data.**

A machine learning platform at the centre of the system makes use of time-series forecasting methods to forecast air quality conditions. Additionally, it uses anomaly detection algorithms to identify anomalous trends or abrupt spikes in pollutant concentrations, which may indicate operational issues or environmental dangers. Environmental agencies and public health experts can take timely action and make informed judgements thanks to these skills.

In addition, the system has an easy-to-use interface that provides both specialists and non-experts with real-time updates and visual representations of air quality statistics. Using IoT technology increases the system's expandability to larger regions and guarantees effective data flow. By providing consistent and accurate air quality information, this technique not only supports urgent mitigation measures but also facilitates strategic environmental and public health planning. Predictive analytics and proactive surveillance are intended to improve environmental governance and public health.

Keywords—LSTM, Linear Regression and Minmax scalar.

1. Introduction:

In developing countries like India, rapid urbanisation and economic expansion have created serious environmental concerns, especially about air pollution, which has detrimental effects on human health as well as a host of ecological issues. Public awareness of these issues is growing, as seen by the rise in respiratory illnesses like asthma, the frequency of acid rain, and the implications for global warming. Accurate air quality forecasting systems are desperately needed, as the relationship between air quality and health is well established. These kinds of technologies are essential for lessening the negative effects that pollution has on the environment and on individuals. For the sake of environmental sustainability and human health, air quality projections must be more accurate. This emphasis is essential for dealing with and lessening the harmful consequences of air pollution, particularly in densely populated urban areas.

2. Related work:

This study investigates the use of machine learning techniques to forecast India's Air Quality Index (AQI). The Air Quality Index (AQI) is a commonly used metric to assess the quality of the air, considering the concentration levels of different pollutants like SO2, NO2, CO2, rspm, and spm. A model that approaches the prediction of AQI as a gradient descent boosted multivariable regression problem was created by the study's student participants. The model forecasts the quality of the air for a given future period by using past pollution data. By using cost estimating approaches that are appropriate for prediction issues, the accuracy of the model was improved, and it was able to anticipate the AQI for vast areas, ranging from counties to entire states, using past pollution data.

Furthermore, the study suggests a composite model to evaluate air pollution levels at different locations in Mumbai and Navi Mumbai using Artificial Neural Networks (ANN) and Kriging. This model makes use of meteorological data and historical information from the neighbourhood Pollution Control Board. With R being used for Kriging and MATLAB for ANN, this model's practical application and validation produced encouraging results in terms of localised air pollution level prediction.

The study also looks at using linear regression and a Multilayer Perceptron (ANN) methodology to forecast the pollution levels for the following day. Using in-depth time series analysis, this method not only predicts impending pollution events but also examines the underlying causes of variations in air quality, improving the forecasting accuracy for subsequent pollution events.

The study's emphasis on tiny particulate pollution is another important feature (PM2.5). The study presents a dual-objective approach that aims to forecast future PM2.5 levels on specified dates in addition to determining PM2.5 concentrations based on particular air quality measurements. Autoregression is used to predict future PM2.5 concentrations from historical data, while logistic regression is used to determine whether a data sample is within contamination standards. The goal of this approach is to offer a dependable way to forecast urban air pollution levels, with a special emphasis on PM2.5.

The paper's main goal is to give a thorough review of the many big data and machine learning techniques used in air quality evaluation and prediction. It contains a case study that shows the application of complex models including the Genetic Algorithm ANN Model, Random Forest, Decision Tree, and Deep Belief Network using data from Shenzhen, China. The advantages and disadvantages of each approach are examined, emphasising how these methods could support air quality control plans that are more precise and flexible.

Overall, this study highlights the promise of sophisticated predictive models in environmental science, especially about their timely and accurate predictions' capacity to manage and mitigate air pollution. Environmental authorities can use it as a useful tool to improve air quality forecasts and put into practice efficient pollution management strategies, which will ultimately lead to better public health outcomes and ecological circumstances.

3. DataSet:

Dataset Description:
About 450,000 entries from different Indian states make up the dataset used in this study, with 60,383 of those records coming from Maharashtra. There are 13 different attributes in this sample of data, including:

1. **Sampling Date:** The precise day that the data on air quality was taken.
2.**Station Code:** An exclusive number assigned to every observation station.
3.**State:** The Indian state in which the observation station is situated.
4. **Location:** The state or city where the data was gathered.
5. **Organisation:** The body or department in charge of gathering data.
6.**Type:** Categorises the monitoring location's surroundings, such as residential or industrial.
7. **SO2 (Sulphur Dioxide):** This important pollutant is measured for its concentration in the atmosphere.
8. **Nitrogen Dioxide (NO2):** This dangerous pollutant's levels are recorded.

9. **Respirable Suspended Particulate Matter (RSPM):** This term describes airborne particles that are tiny enough to enter the respiratory system through inhalation.
10. **Suspended Particulate Matter, or SPM:** measures a wider range of airborne particles.
11.**Monitoring Station Location:** Specifies where the air monitoring apparatus is situated.
12. **PM2.5:** Monitors tiny particulate matter, which is a crucial air quality indicator, having a diameter of less than 2.5 micrometres.
13.**Date:** An improved sampling date representation for a more understandable analysis.

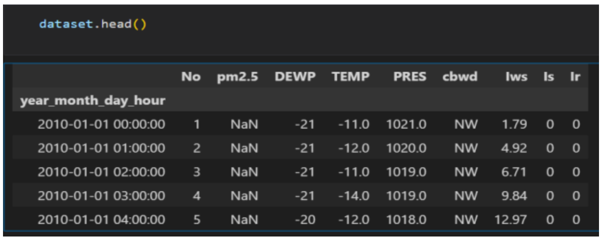
The PM2.5 column of the dataset, which is essential for assessing air quality, sadly has many null values, making analysis difficult. 20% of the data was set aside for accuracy testing and the remaining 80% was used to train the predictive models for the study. A thorough examination of the trends in Maharashtra's air quality and the variables affecting pollution levels is made possible by this methodical methodology.

Feature selection and preprocessing:

We only used data from Maharashtra in our investigation, yielding a dataset with 60,383 items. With this particular focus, the state-indicating column was quickly removed because it was no longer needed. The 'pm2_5' column was likewise disregarded because all of its entries were null, offering no useful information. The agency's name, which has no bearing on pollution levels, and the station code, which didn't add anything to our research, were also eliminated.
As the 'date' field provides a more accurate timestamp and prevents data redundancy, we removed the 'sampling date' from the dataset to further streamline it. Furthermore, the monitoring station's location was eliminated because it was thought to be unrelated to our goals.
Once the dataset was cleaned, it was further adjusted to include the categories 'residential', 'industrial', and 'other' in the 'type' field, which made the analysis easier. To preserve data integrity, missing values for SO2 and NO2 were substituted with their corresponding means. The 'date' field only contained three missing items, so those were easily eliminated.

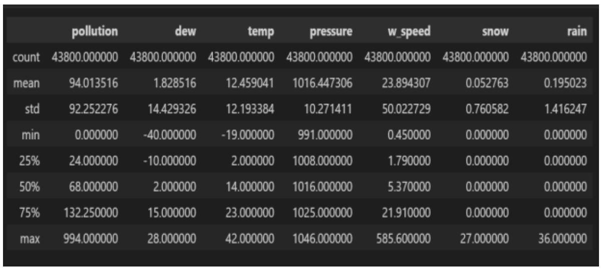
After preprocessing, our dataset has 60,380 rows and seven columns, which allows us to efficiently focus on the most important variables for our investigation of the trends in air quality and the factors that contribute to it in Maharashtra.

Data



```
dataset.head()
```

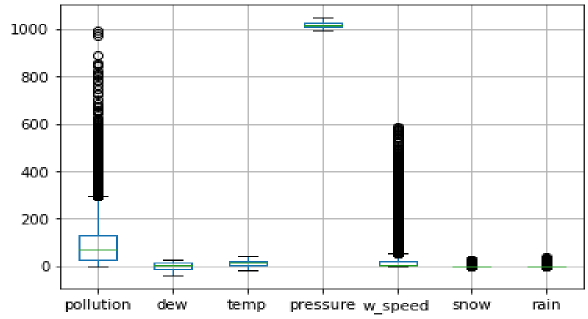
```
                     No  pm2.5  DEWP  TEMP    PRES  cbwd    Iws  Is  Ir
year_month_day_hour
2010-01-01 00:00:00   1    NaN   -21  -11.0  1021.0   NW   1.79   0   0
2010-01-01 01:00:00   2    NaN   -21  -12.0  1020.0   NW   4.92   0   0
2010-01-01 02:00:00   3    NaN   -21  -11.0  1019.0   NW   6.71   0   0
2010-01-01 03:00:00   4    NaN   -21  -14.0  1019.0   NW   9.84   0   0
2010-01-01 04:00:00   5    NaN   -20  -12.0  1018.0   NW  12.97   0   0
```



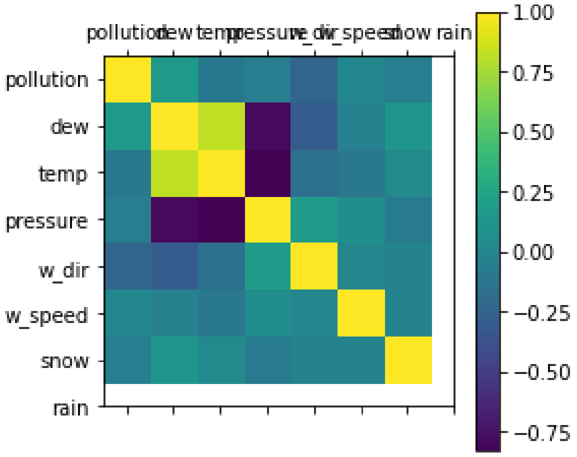
```
          pollution           dew          temp      pressure       w_speed          snow          rain
count  43800.000000  43800.000000  43800.000000  43800.000000  43800.000000  43800.000000  43800.000000
mean      94.013516      1.828516     12.459041   1016.447306     23.894307      0.052763      0.195023
std       92.252276     14.429326     12.193384     10.271411     50.022729      0.760582      1.416247
min        0.000000    -40.000000    -19.000000    991.000000      0.450000      0.000000      0.000000
25%       24.000000    -10.000000      2.000000   1008.000000      1.790000      0.000000      0.000000
50%       68.000000      2.000000     14.000000   1016.000000      5.370000      0.000000      0.000000
75%      132.250000     15.000000     23.000000   1025.000000     21.910000      0.000000      0.000000
max      994.000000     28.000000     42.000000   1046.000000    585.600000     27.000000     36.000000
```

4. Exploratory Data Analysis:

Box plot:



CORRELATION MATRIX:



Forecasting air pollution is essential for minimising environmental harm and safeguarding public health. Researchers can identify the primary determinants of air quality by looking through historical data and generating a correlation matrix. The development of accurate prediction models made possible by this study enables proactive approaches and well-informed policymaking. Comprehending these intricate relationships is essential for effectively managing pollution and advancing sustainable urban development.

5. Results and Discussion:

We can ascertain the future data points with the use of time series analysis.
Models used in the same situation include:

1) The test's MSE for the autoregressive model, or AR model, is 166.358.
One kind of time series model is autoregression, which anticipates the value at the subsequent time step by feeding observations from previous time steps into a regression equation.
It's a rather simple idea that can yield accurate forecasts for a range of time series problems.
What that equals is b0 + b1 * X1.

where yhat is the forecast, X is an input value, and the coefficients b0 and b1 are found by optimising the model with training data.

This technique can be used with time series in which the observations at previous time steps serve as the input variables, also known as lag variables.

For example, we can anticipate the value for the next time step (t+1) based on the observations from the previous two-time steps (t-1 and t-2). This would look somewhat like this if it were a regression model:

X(t+1) = b0 + b1*X(t-1) + b2*X(t-2).

The regression model is called an autoregression (regression of self) because it makes use of data from the same input variable at earlier time steps.

6.Conclusion:

Our research emphasises how urgently cities like Pune and Mumbai, where SO2 levels are alarmingly rising, need to act. To efficiently estimate SO2 concentrations, we employed Autoregressive (AR) and Autoregressive Integrated Moving Average (ARIMA) models. We streamlined our model by eliminating superfluous parameters such the location_monitoring_station or station code that did not improve the predicted accuracy.

The safety levels for SO2 are now set at 0.20 parts per million on an hourly average, 0.08 parts per thousand over a 24-hour period, and 0.02 parts per million on an annual basis. We proactively handle any health hazards by monitoring these concentrations to make sure they stay within acceptable limits thanks to our predictive efforts.

A crucial component of our study also examined PM2.5, a particulate matter that has serious consequences for cardiovascular and respiratory health, including illnesses like asthma and bronchitis as well as abnormalities in the heart like heart attacks. Because PM2.5 has such a large negative influence on health, it is imperative that we keep tracking and predicting it.

The disorganised data according to the date column posed a significant barrier to our study and reduced the efficacy of our models. This problem brought to light the significance of city-specific forecasts relative to state-wide projections since local environmental variables vary. To improve our study and offer useful insights, we intend to compute the Air Quality Index (AQI) and use categorization methods in the future.

As a result, our study not only highlights the urgent need for action in severely polluted places, but it also makes recommendations for future research topics, including improved PM2.5 surveillance and sophisticated AQI modelling. Such programmes are essential for creating strategies that effectively reduce pollution and safeguard public health.

REFERENCES

[1] Mrs. A. GnanaSoundariMtech, (Phd) ,Mrs. J. GnanaJeslin M.E, (Phd), Akshaya A.C. “Indian Air Quality Prediction And Analysis Using Machine Learning”. International Journal of Applied Engineering Research ISSN 0973-4562 Volume 14, Number 11, 2019 (Special Issue)

[2] Suhasini V. Kottur , Dr. S. S. Mantha. “An Integrated Model Using Artificial Neural Network(Ann) And Kriging For Forecasting Air Pollutants Using Meteorological Data”.International Journal of Advanced Research in Computer and Communication Engineering ISSN (Online) : 2278-1021 ISSN (Print) : 2319-5940 Vol. 4, Issue 1, January 2015

[3] RuchiRaturi, Dr. J.R. Prasad .“Recognition Of Future Air Quality Index Using Artificial Neural Network”.International Research Journal of Engineering and Technology (IRJET) .e-ISSN: 2395-0056 p-ISSN: 2395-0072 Volume: 05 Issue: 03 Mar-2018

[4] Aditya C R, Chandana R Deshmukh, Nayana D K, Praveen Gandhi Vidyavastu .” Detection and Prediction of Air Pollution using Machine Learning Models”. International Journal of Engineering Trends and Technology (IJETT) – volume 59 Issue 4 – May 2018

[5] Gaganjot Kaur Kang, Jerry ZeyuGao, Sen Chiao, Shengqiang Lu, and Gang Xie.” Air Quality Prediction: Big Data and Machine Learning Approaches”. International Journal of Environmental Science and Development, Vol. 9, No. 1, January 2018

[6] https://machinelearningmastery.com/autoregressionmodels-time-series-forecasting-python/

[7] https://machinelearningmastery.com/arima-fortime-series-forecasting-with-py