IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

AIR QUALITY INDEX PREDICTION AND CLASSIFICATION

Ardra Nandakumar, Jasmine Jose
MSc Scholar, Assistant Professor
Department of Computer Science,
St. Joseph's College (Autonomous), Irinjalakuda, Thrissur, India

Abstract: Air quality prediction is one of the complex tasks due to its dynamic nature, volatility, and variability in the time and space of pollutants. At the same time, being able to model, predict, and monitor the quality of air is becoming more relevant due to the observed critical impact of air pollution on the health of citizens and the environment. This paper aims to predict the monthly Air Quality Indices using ARIMA (Autoregressive Integrated Moving Average) model. Then categorize AQI into different classes. This knowledge will help us to reduce its level to a lesser than harmful range. The prediction and categorization will be based on a data set of atmospheric conditions in a specific area.

Index Terms - AQI, ARIMA MODEL, Stationary Series.

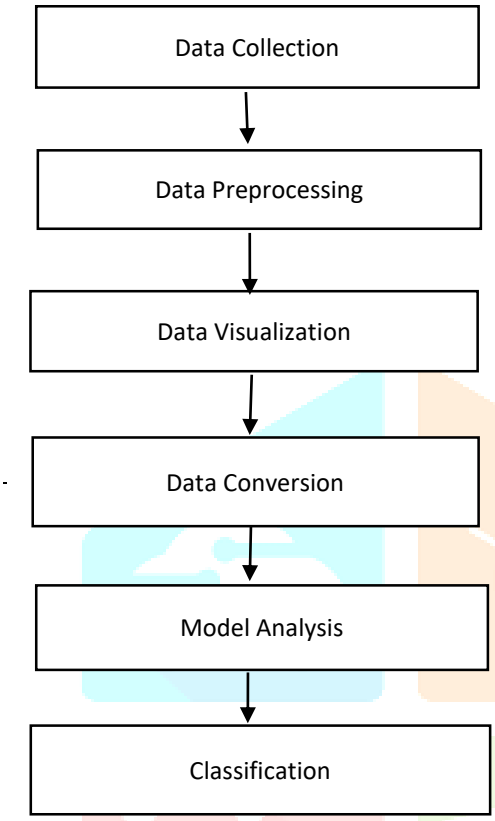
I. INTRODUCTION

Air pollution is one of the greatest threats we are facing today. It is rapidly increasing due to various human activities, and the occurrence of particulates, chemicals, or biological resources into the environment that cause unexpected, humans' death, or disease, damage source of revenue, or spoil the natural environment. The direct impact of air pollution problems on human health resulted in an increased public awareness of the air quality in developing and developed countries. The commonly occurring pollutants include carbon monoxide (CO_2), Respirable Suspended Particulate Matter (RSPM), Sulphur dioxide (SO_2), Nitrogen dioxide (NO_2), Ozone etc. The increase in the concentration of pollutants in the air will be life-threatening and will cause several diseases. So, it is necessary to monitor and forecast the criteria pollutants in the air.

The prediction of air pollutants can be made through models. In most of the air quality predictions, Gaussian dispersion models are used. These dispersion models have some physical basis. So, information about the source of pollutants and other parameters are not known to us. Statistical models can be used to overcome these difficulties. They are capable of forecasting the concentration of pollutants in air. The AQI is defined as an index or rating scale for reporting the daily combined effect of ambient air pollutants recorded in the monitoring sites. A regression model is also used for predicting air pollution. A study is made for daily air quality forecasting of air pollutants through ARIMA and Multiple Linear Regression (MLR) models. The main objective of the present study is to develop forecasting models for predicting the monthly Air Quality Indices, which can provide timely information to the public to take precautionary measures to protect their health.

II. METHODOLOGY

The proposed system for prediction of future AQI and classification involves the following steps:



1) DATA COLLECTION

The action of collecting data required to calculate the Air Quality Index. Air Quality Index is calculated by using data as a measure of air pollutants. These data are collected from Pollution Control Board (PCB). PCB collects this data daily using a hardware device in a different region. We collect a monthly average of this data. Different air pollutants contribute to different AQI. The Highest AQI value is considered as AQI in that region. In the proposed system, we collect measures of classical pollutants like NO_2, SO_2, and RSPM monthly average of 5 years from PCB. We have to train and predict month-wise NO_2, SO_2, and RSPM based on these data and calculate AQI.

2) DATA PREPROCESSING

As part of preprocessing, the missing values must be removed from the data-set to ensure that the results generated are more accurate. Since most data is available in numerical format, it is possible to calculate the mean, median, or mode of the features and replace them with missing values. This approximation of values can be used to add variance to the data. To handle the missing values, the above three approximations have to be replaced. After there are eight readings for each day, it must be normalized in such a way that for the particular day only one reading is obtained.

3) DATA VISUALIZATION

Data visualization provides additional interpretation possibilities. Visual presentation for individuals is often more readable than any other presentation of data weather data in itself is massive. Since the data is in enormous amounts, data visualization helps to understand the dataset. It helps to see patterns, trends, and correlations in the values of attributes that may go unnoticed. Climate data patterns and their variations during the year can be identified using this.

4) DATA CONVERSION

Air pollutant data may be in non-stationary time series with statistical properties that change over time. Prior to the predictive modelling, it is necessary to make sure that these statistical properties are constant. A time series in which the statistical properties such as mean, variance and auto-correlation are constant over time is known as stationary time series. Conversion to stationary time series using:

a. **Differencing**: The difference between the successive terms is calculated. It is performed to avoid the problem of the varying mean. In the process of fitting an ARIMA model, it is necessary to make a stationary time series through differencing. Mathematically, differencing can be written as:
YT'=YT – Y(T-1), YT is the value at a time T.
b. **Transformation**: To stabilize a series of non-constant variance, transformations are used. Power transformation, square root transformation, and log transformation are the different methods used for transformation.

5) MODEL ANALYSIS

AQI prediction is the system's main objective. ARIMA model can be used for this purpose, as it is considered as the most suitable statistical method. ARIMA is the short form for Auto-Regressive Integrated Moving Average. It is a class of models that will describe a given time series on the basis of its previous values.
ARIMA has 3 components:
a. **AR**: Auto-regressive. This means that the variable is regressed on its own lagged values.
b. **I:** Integrated. This means that the data values should be substituted by the difference between the values and their previous values.
c. **MA**: Moving Average. This means that the regression error is a linear union of terms whose values occurred concurrently and at various times in the past.

The aim of each of these components is to shape the model to fit the data as well as possible. Time series consists of observations on a variable over time. Such information must be gathered daily (e.g., weather), weekly (e.g., monthly supply), or annually (e.g., government budget) at periodic intervals. Time Series is used in statistics, finance, prediction of earthquakes, forecasting of weather, and many other applications.

STATIONARY SERIES

For a series to be stationary, there are three conditions to be satisfied:
a. The mean of the series should not be a constant. It should be a function of time.
b. Homoscedasticity: The variance of the series should not be a function of time.
c. The covariance between i^{th} term and the $(i + m)^{th}$ term of the series should not be a function of time.

ARIMA MODEL

ARIMA stands for Auto-Regressive Integrated Moving Averages. For a stationary time series, the ARIMA forecasting is a linear equation.
Number of AR (Auto-Regressive) terms (p): AR terms are just lagging of dependent variable. For example, if p is 6, the predictors for x(t) will be x(t-1) ….x(t-6).
Number of MA (Moving Average) terms (q): MA terms are lagged forecast errors in the prediction equation. For example, if q is 6, the predictors for x(t) will be e(t-1) …. e(t-6) where e(i) is the difference between the moving average at i^{th} instant and actual value.

5) CLASSIFICATION

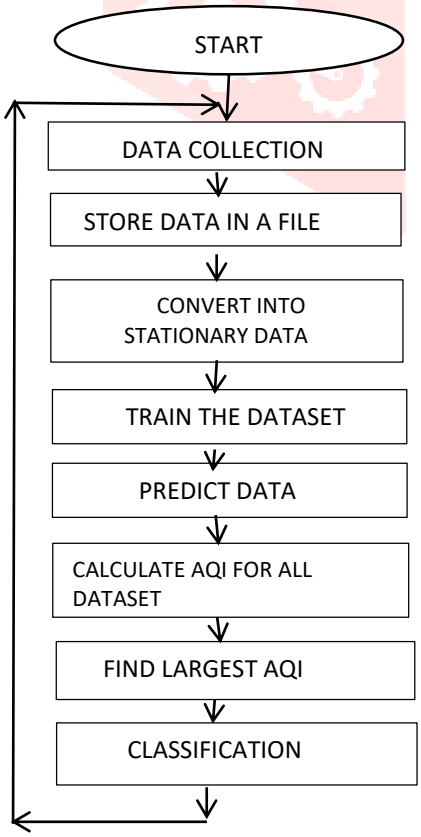
AQI is of each air pollutant is calculated using the equation:

$I=I_h-I_lC_h-C_lC-C_l+I_l$
Here,
I=the Air Quality Index
I_h= the index break-point corresponding to C_h
I_l=the index break-point corresponding to C_l
C_h=the concentration breaking point that is >=C
C_l=the concentration breaking point that is <=C
C = the pollutant concentration
After calculation and prediction largest AQI value is considered as the AQI of that region. Then this AQI is classified in a different class as good, moderate, and bad based on a different breaking point.

III. PROPOSED SYSTEM

The proposed system entitled "Air Quality Index Prediction and Classification" aimed to predict AQI of future month-wise and classify them in different categories. Here we collect five-year monthly concentrations of NO_2, SO_2, and RSPM from the Pollution Control Board. It consists of monthly minimum, average, and maximum value air pollutants on that month. And calculate the AQI of each pollutant and assign AQI as the largest AQI value. Train these data and predict future value. Then categorize AQI into a different class. This helps common people to know about air quality status.

IV. ARCHITECTURE DIAGRAM



V. ACKNOWLEDGEMENT

I am extremely thankful to Principal Dr. Sr. Anis K.V, Head of the Department Sr. Siji P.D for giving me constant support and guidance for preparing and presenting the paper. I express my sincere gratitude towards my guide, Ms. Jasmine Jose for her help, encouragement and inspiration during the preparation of this paper.

VI. CONCLUSION

The reduction of air pollutant levels is one of the most important tasks for protecting both human health and the environment. People must know what the level of pollution in their surroundings is and takes a step towards fighting against it. The conclusions from various studies show that the use of machine learning models such as logistic regression and autoregression helps to detect the quality of air and predict the level of AQI in the future and classify them into different classes. The proposed system will help common people as well as those in the meteorological department to detect and predict pollution levels and take the necessary action following that. Also, this will help the public set up a data source for small localities which are usually left out in comparison to the large cities.

REFERENCES

[1] Chelani AB, Rao CVC, Phadke KM, Hasan MZ. Prediction of Sulphur dioxide concentration using artificial neural networks. Environ Modell Softw 2002; 17:161–

[2] Finzi G, Tebaldi G. A mathematical model for air pollution forecast and alarm in an urban area. Atmos Environ 1982;16(9):2055–9.

[3] Aditya CR, Chandana R Deshmukh, Nayana D K, Praveen Gandhi Vidyavastu; "Detection and Prediction of Air Pollution using Machine Learning Models", International Journal of Engineering Trend and Technology (IJETT) Volume 59 Issue 4-May 2018

[4] Macro Castelli, Fabiana Martin Clemente, Ales Popovic, Sara Silve, Leonardo Vanneshi; "A Machine Learning Approach to Predict Air Quality in California", Complexity, vol.2020

[5] Balram Ambade, Sampad Ghosh and Basant Shubhankar; "Ambient Air Quality of Jamshedpur city: A Study with reference to SO_2, NO_2, RSPM and SPM contents", Journal of Applicable Chemistry,2013,2(3):539-544

[6] Inderjeet Kaushik, Rinki Melwani; "TIME SERIES ANALYSIS OF AMBIENT AIR QUALITY AT ITO INTERSECTION IN DELHI (INDIA)", Journal of Environmental Research and Development, vol.2 No.2, October-December 2007.

[7] Yun-Chia Liang, Yona Maimury, Angela Hsiang-Ling Chen and Josue Rodolfo Cuevas Juarez; "Machine Learning-Based prediction of Air Quality ", applied sciences,21 December 2020.

[8] Snehal Sirsikar, Priya Karemore; "Review paper on Air Pollution Monitoring System", International Journal of Advanced Research in Computer and Communication Engineering, Vol 4, issue 1, January 2015.