Project Report

Project Name : Time Series Analysis using ARIMA model for Air Pollution Prediction
Team Members: Akshith Reddy Chada (RH34539), Pooja Gopu (SR21814)

Introduction of the project :

We all know that air quality has become a major public health issue, particularly in India, where the majority of the population suffers from poor air quality. There are several chemicals in the atmosphere that produce air pollution, resulting in a hazardous environment. But we don't know in what amounts they are present. According to the World Health Organization, nearly 80% of the population lives in metropolitan areas where air quality is sensitive to the essential quality requirements that exceed WHO limits. It claims that air pollution causes around 4.2 million fatalities per year and that at least 140 million people breathe bad air quality that is ten times worse than the WHO's safety standard. Thirteen Indian cities out of twenty throughout the world have the highest yearly levels of air pollution. In order to bring awareness among people and government, based on the values of the pollutants in the previous years, we have tried to predict the pollutants values for the future years. We used the ARIMA model for the prediction of the pollutants. Previously we have seen many models using LSTM and also other neural networks for time series analysis, comparing all those models with ARIMA, we found that ARIMA model was more accurate in predicting the values.

The dataset used in this model was from the website below, this dataset was released by the Ministry of Environment and Forests and Central Pollution Control Board of India under the National Data Sharing and Accessibility Policy (NDSAP).
https://data.gov.in/catalog/historical-daily-ambient-air-quality-data
In order to apply time series analysis we require only two features i.e date and values of a particular pollutant. The proposed model is used to segregate the data belonging to a specific type of the area which is under a particular location in a state. In ARIMA, we try to feed the model with the historical data and by using that data, the model gets trained (we check the stationarity of the model, and turn the model into a stationary one) and it tries to predict the pollutants for the future years.

Contribution of each team member:

Akshith: Tried to use different datasets in the beginning, got the appropriate dataset from the Central Government of India as mentioned above, worked with Data Preprocessing, struggled a bit with data cleaning. Calculated various error rates and comparisons were made to check if the model is accurate or not.

Pooja: Model Analysis was done by me, initially faced issues while trying to forecast the model, but with a lot of experiments done, we could resolve those issues. Tried hard to understand the differencing part of the model, which was represented statistically all over the internet. I was also responsible for the calculated results.

Mechanism of the model:

In order to apply time series analysis we require only two features i.e date and values of a particular pollutant. The proposed model is used to segregate the data belonging to a specific type of the area which is under a particular location in a state. The values of a certain pollutant recorded in such a type of location and the observation date are the only two features which are used in our model. In the dataset, if there are any missing values or incorrect data they will be replaced with the correct data. In data cleaning unnecessary data features are removed, formats are fixed and missing values are filled.

In order to apply the ARIMA model, the model should be stationary. For a model to be stationary it should have constant mean, constant variance and autocorrelation that does not depend on time. There are two types of tests to check whether a model is stationary or not. They are Rolling Statistics (RS) and Augmented Dickey Fuller Test (ADFT). In Rolling Statistics, plot the moving average or moving variance and observe if it varies with time. If it is yes, the model is a non stationary one. Or else it is a stationary model. In the Augmented Dickey Fuller Test, the Null hypothesis is that the time series is non-stationary. The results of the dickey fuller test indicate whether the model is stationary or not. If the obtained p-value is very less and if the critical values are greater than test statistic then the model is said to be stationary. After performing the tests for stationarity, if the model is not stationary then transform the time series into stationary by using differencing. Differencing is a process of transforming non stationary time series to stationary time series. The first differencing value is the difference between the current time period and the previous time period. If the model is still not stationary then go for second differencing. Repeat this until a stationary model is obtained. ARIMA model is a combination of Auto Regressive (AR) model and Moving Average (MA) model with differencing part (I). Auto Regressive is the correlation between the previous time period(s) and the current time period. If there is any such correlation, then it is an auto regressive model.
Let Xt be a function of the lags. $Xt = \alpha + \beta1 * Xt-1 + \beta 2 * Xt-2 + ... + \beta a * Xt-a$
There is always some kind of noise or irregularity attached in the time series. There is a need to average that noise in order to smoothen it and hence we use the Moving average model.
$Xt = \alpha + t + \varphi1 * t-1 + \varphi2 * t-2 + ... + \varphi b * t-b$ where Xt depends only on the lagged forecast errors.

Combining the above two equations we get, $Xt = \mu + \beta1 * Xt-1 + ... + \beta a * Xt-a + \varphi1 * t-1 + ... + \varphi b * t-b$ (3) where μ is the constant term , $\beta1Xt-1 + ... + \beta aXt-a$ are the AR terms and $\varphi1t-1 + ... + \varphi bt-b$ are the MA terms.

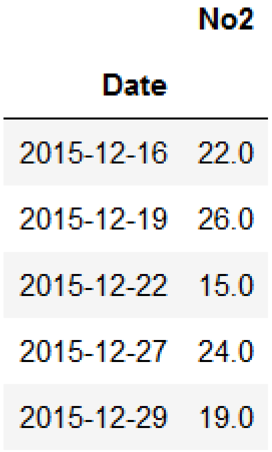
ARIMA model has three main parameters: p,d,q where, p: number of auto regressive lags d: number of differencing required to make the time series stationary q: order of moving average which specifies how deviations from the time series mean for previous values are used to predict current values.

We tried to divide the data set randomly for training(70%) and testing(30%). Initially the training model was not stationary, then we did differencing. Then again we checked if our model was stationary or not, and this time it resulted in a stationary one. We then selected parameters p,d,q by using PACF and ACF graphs. The value at which the graph touches the zero line for the first time gives the value of p in the PACF graph and the value of q in the ACF graph. Hence we predicted values for the training set and evaluated the training and testing data sets.

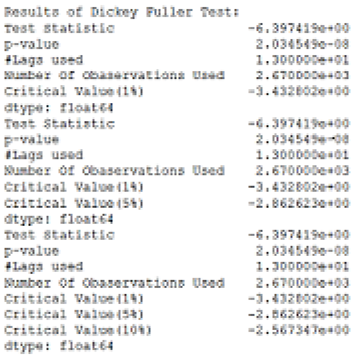
Demonstration of our team's work:

We have taken the input values of NO2 at areas belonging to Residential places in a city in India and also noted the date on which the observations are recorded. The stationarity of the time series is checked with the help of RS and ADFT. In ADFT for a stationary model p value should be very less and critical value should be greater than the test statistic.

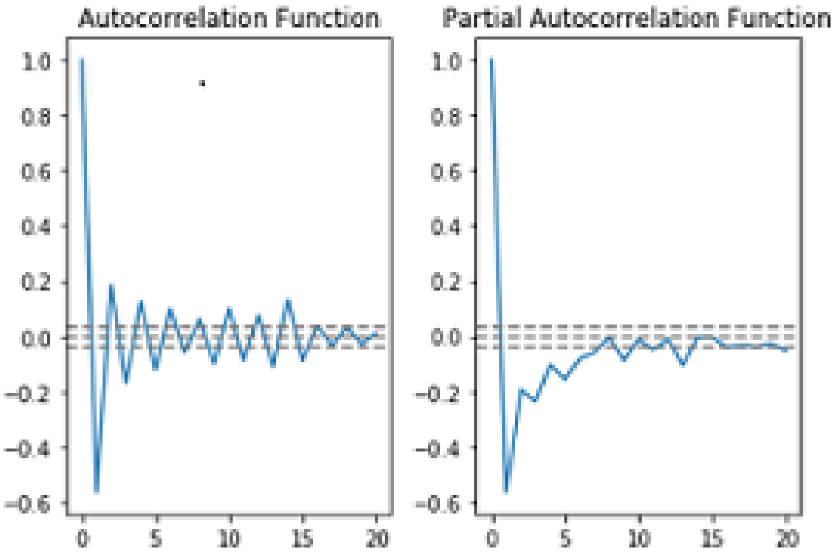
When this is performed p-value is obtained as 0.000970 but the test statistic is greater than the critical value resulting in a non-stationary model. Furthermore, differencing is done and again the stationarity test is repeated which proves the model to be stationary. Thus the value of d is 1.



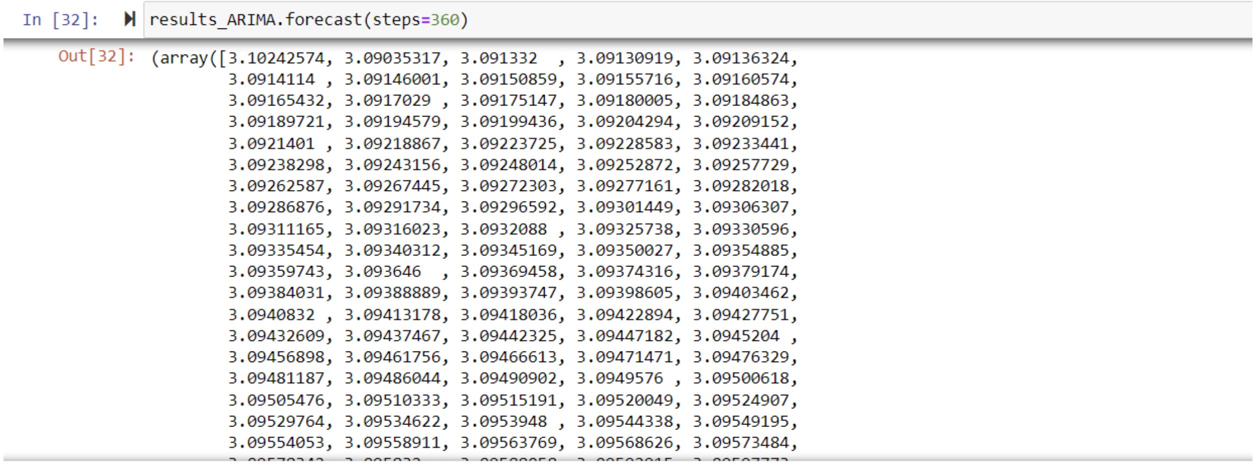
	No2
Date	
2015-12-16	22.0
2015-12-19	26.0
2015-12-22	15.0
2015-12-27	24.0
2015-12-29	19.0



```
Results of Dickey Fuller Test:
Test Statistic                  -6.397419e+00
p-value                          2.034549e-08
#lags used                       1.300000e+01
Number Of Obaservations Used     2.670000e+03
Critical Value(1%)              -3.432802e+00
dtype: float64
Test Statistic                  -6.397419e+00
p-value                          2.034549e-08
#lags used                       1.300000e+01
Number Of Obaservations Used     2.670000e+03
Critical Value(1%)              -3.432802e+00
Critical Value(5%)              -2.862623e+00
dtype: float64
Test Statistic                  -6.397419e+00
p-value                          2.034549e-08
#lags used                       1.300000e+01
Number Of Obaservations Used     2.670000e+03
Critical Value(1%)              -3.432802e+00
Critical Value(5%)              -2.862623e+00
Critical Value(10%)             -2.567347e+00
dtype: float64
```



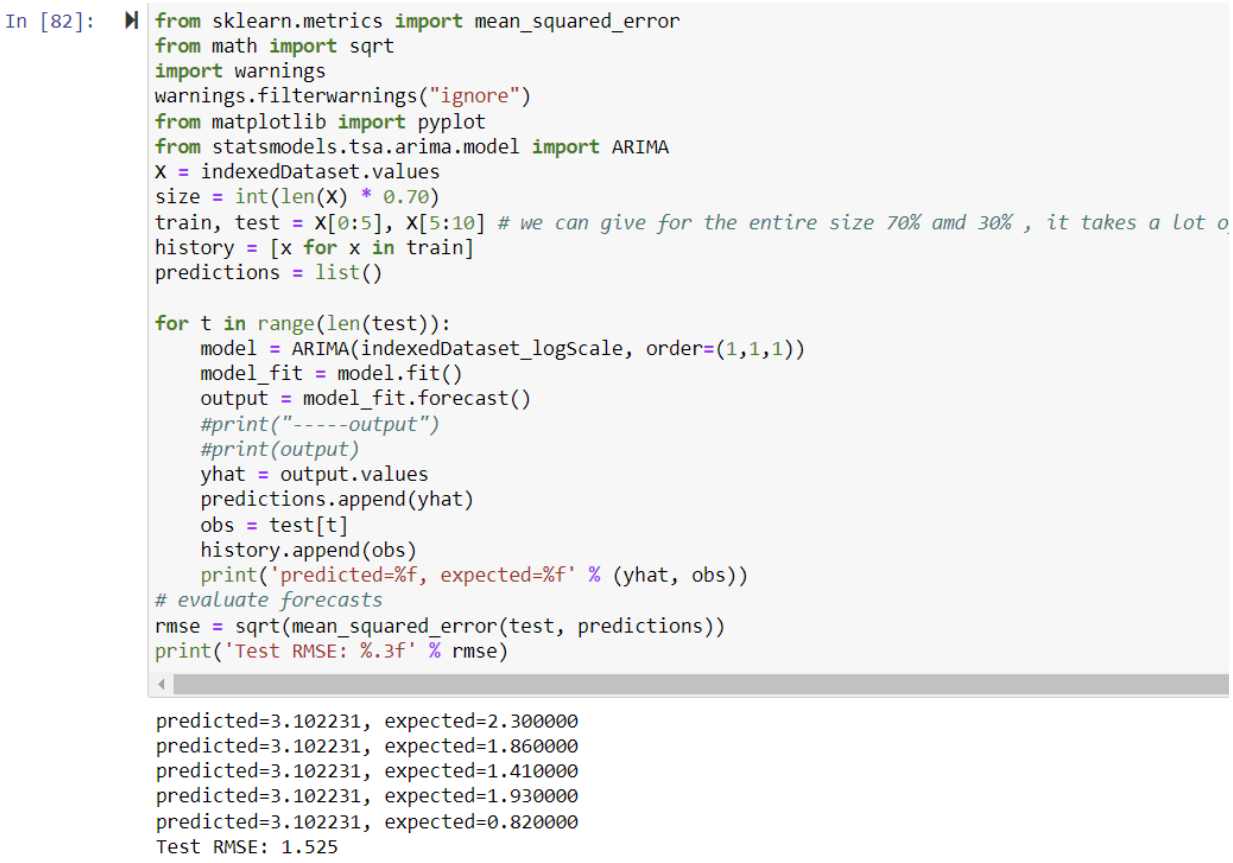
From the PACF graph, the value at which the graph touches the zero line for the first time is ‘1’.Hence the value of p is 1. From the ACF graph, the value at which the graph touches the zero line for the first time is ‘1’. Therefore the value of q is also 1. Now the parameters (p,d,q) for the ARIMA model are (1,1,1). Then the graph is plotted by taking the produced ARIMA results. In this way the values of NO2 are forecasted for the next 10 years.



```
In [32]:    results_ARIMA.forecast(steps=360)

Out[32]: (array([3.10242574, 3.09035317, 3.091332  , 3.09130919, 3.09136324,
                 3.0914114 , 3.09146001, 3.09150859, 3.09155716, 3.09160574,
                 3.09165432, 3.0917029 , 3.09175147, 3.09180005, 3.09184863,
                 3.09189721, 3.09194579, 3.09199436, 3.09204294, 3.09209152,
                 3.0921401 , 3.09218867, 3.09223725, 3.09228583, 3.09233441,
                 3.09238298, 3.09243156, 3.09248014, 3.09252872, 3.09257729,
                 3.09262587, 3.09267445, 3.09272303, 3.09277161, 3.09282018,
                 3.09286876, 3.09291734, 3.09296592, 3.09301449, 3.09306307,
                 3.09311165, 3.09316023, 3.0932088 , 3.09325738, 3.09330596,
                 3.09335454, 3.09340312, 3.09345169, 3.09350027, 3.09354885,
                 3.09359743, 3.093646  , 3.09369458, 3.09374316, 3.09379174,
                 3.09384031, 3.09388889, 3.09393747, 3.09398605, 3.09403462,
                 3.0940832 , 3.09413178, 3.09418036, 3.09422894, 3.09427751,
                 3.09432609, 3.09437467, 3.09442325, 3.09447182, 3.0945204 ,
                 3.09456898, 3.09461756, 3.09466613, 3.09471471, 3.09476329,
                 3.09481187, 3.09486044, 3.09490902, 3.0949576 , 3.09500618,
                 3.09505476, 3.09510333, 3.09515191, 3.09520049, 3.09524907,
                 3.09529764, 3.09534622, 3.0953948 , 3.09544338, 3.09549195,
                 3.09554053, 3.09558911, 3.09563769, 3.09568626, 3.09573484,
```

When trying to calculate the errors, the residual sum of the squares(RSS) value turned out to be the least for the ARIMA when compared with the AR model or MA model.



In [82]:

```
from sklearn.metrics import mean_squared_error
from math import sqrt
import warnings
warnings.filterwarnings("ignore")
from matplotlib import pyplot
from statsmodels.tsa.arima.model import ARIMA
X = indexedDataset.values
size = int(len(X) * 0.70)
train, test = X[0:5], X[5:10] # we can give for the entire size 70% amd 30% , it takes a lot o
history = [x for x in train]
predictions = list()

for t in range(len(test)):
    model = ARIMA(indexedDataset_logScale, order=(1,1,1))
    model_fit = model.fit()
    output = model_fit.forecast()
    #print("-----output")
    #print(output)
    yhat = output.values
    predictions.append(yhat)
    obs = test[t]
    history.append(obs)
    print('predicted=%f, expected=%f' % (yhat, obs))
# evaluate forecasts
rmse = sqrt(mean_squared_error(test, predictions))
print('Test RMSE: %.3f' % rmse)
```

```
predicted=3.102231, expected=2.300000
predicted=3.102231, expected=1.860000
predicted=3.102231, expected=1.410000
predicted=3.102231, expected=1.930000
predicted=3.102231, expected=0.820000
Test RMSE: 1.525
```

The accuracy calculated would be 98.47%

References:

1. Singh, Pratyush, T. Lakshmi Narasimhan, and Chandra Shekar Lakshminarayanan.: DeepAir: Air Quality Prediction using Deep Neural Network. TENCON IEEE Region Conference (TENCON). IEEE,(2020)
2. Bhardwaj, R., and D. Pruthi.: Development of model for sustainable nitrogen dioxide prediction using neuronal networks. International Journal of Environmental Science and Technology (2020) 1-10 11.
3. Shah, Jalpa, and Biswajit Mishra: Analytical Equations based Prediction Approach for PM2. 5 using Artificial Neural Network (2020)
4. Saha, D., Soni, K., Mohanan, M. N., Singh, M.: Long-term trend of ventilation coefficient over Delhi and its potential impacts on air quality. Remote Sensing Applications: Society and Environment, (2019) 15, 100234.
5. Geetha, A., and G. M. Nasira.: Time series modeling and forecasting: Tropical cyclone prediction using ARIMA model. 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom). IEEE,(2016)
6. Bose, R., Dey, R. K., Roy, S., Sarddar, D.: Time Series Forecasting Using Double Exponential Smoothing for Predicting the Major Ambient Air Pollutants. In Information and Communication Technology for Sustainable Development. Springer, Singapore (2020) 603-613 15.
7. Adebiyi, Ayodele Ariyo, Aderemi Oluyinka Adewumi, and Charles Korede Ayo.: Comparison of ARIMA and artificial neural networks models for stock price prediction (2014)