Jaypee Institute of Information Technology, Noida
MINOR PROJECT - I
Work Summary Sheet

Enroll. No.(s) - 19103175,19103070,19103071
Name of student(s) - Khushi Jain, Megha Agarwal, Pavini Jain
Supervisor name – Dr. Parul Agarwal

PROJECT TITLE:
TIME SERIES FORECASTING OF AIR QUALITY PREDICTION

1. Motivation behind the project

Contamination of the air, in particular in metropolitan areas, is a very well-known problem. The ever-growing population of cities and the increasing level of motorization contribute to the ever-increasing traffic volume, and consequently, the ever-increasing exhaust gases emissions. At the same time, the thickening of city buildings reduces ventilation and increases the porosity of surface, which ends up decreasing the effect of the wind on the evacuation of contamination.

To estimate the impact on our health, we must be aware of the air quality in our neighbourhood, city, and country. The main goal of this project is to investigate the state and quality of the air by measuring the Air Quality Index (AQI).

Therefore, building a forecasting system for predicting the air quality based on the levels of concentration of individual pollutants and various meteorological parameters will be useful for the country's health.

Problem Statement

In this project our main aim is to develop an efficient approach for forecasting the air quality index of Bangalore using meteorological parameters and concentration of major air pollutants using various Machine Learning Algorithms and Deep Neural Networks like LSTM.

2. Type of project

Research cum development project

3. Critical Analysis of research paper

Year of Publication: **2021**

Title: Air Quality Prediction Model Based on Spatiotemporal Data Analysis and Metal earning

*Problem Statement: P*redict the air quality of any monitoring station based on the existing weather and environmental data while considering the spatiotemporal correlation among monitoring stations and maintain the accuracy and stability of the forecast even when the available data is severely insufficient.

Proposed Methodology: They proposed a spatiotemporal model GATLSTM by combining LSTM and GAT for air quality prediction, then design a metal earning algorithm for GAT-LSTM for transfer learning.

Year of Publication: 2020

Title: A Machine Learning Approach to Predict Air Quality in California

Problem Statement: The study aims to build models for hourly air quality forecasting for the state of California, using approaches like a variant of support vector machines. It also aims to forecast pollutant and particulate levels and to predict the air quality index (AQI)

Proposed Methodology: The proposal was to build an SVR model for the prediction of each pollutant and particulate measurement on an hourly basis and an SVR model to predict the hourly air quality index for the state of California. Both PCA SVR-RBF and SVR-RBF achieved similar performance in forecasting the AQI.

Year of Publication: 2019

Title: Air Quality Prediction in Visakhapatnam with LSTM based Recurrent Neural Networks

Problem Statement: To develop an efficient approach for modelling and prediction of air quality

Proposed Methodology: They proposed a LSTM based RNN framework for forecasting concentrations of air pollutants in Visakhapatnam.

Year of Publication: 2018

Title: A Multiple Kernel Learning Approach for Air Quality Prediction

Problem Statement: The aim is to predict the air quality health index (AQHI) in Hong Kong and the PM2.5 individual air quality level (IAQL) in Beijing. AQHI and IAQL are scales designed to help understand the impact of air quality on health.

Proposed Methodology: As the source data is coming from different modalities. Therefore, instead of using just a single kernel, multiple kernels are combined. Multiple kernel learning is conceptually similar to single kernel learning. In other words, single kernel leaning is a special case of MKL. In this model, a novel multiple kernel learning-based approach with SVC as the base learner was proposed for the near future's air quality prediction.

Year of Publication: 2017

Title: Urban air quality forecasting based on multidimensional collaborative SVR

Problem Statement: The aim of this study is to present a new model for AQI forecasting using collaborative multiple city air quality data as input

Proposed Methodology: A Multi-dimensional collaborative SVR model was presented. The RMSE values of the training and the testing datasets are <12 and are almost similar for the training and the testing datasets for most of the cases, hence we can conclude that the support vector regression model is strong and reliable for predicting the AQI values.

Year of Publication: 2016

Title: RAQ–A Random Forest Approach for Predicting Air Quality in Urban Sensing Systems

Problem Statement: This paper uses urban sensing data to solve the problem of air quality inference which means to infer the unknown air quality of areas by using all kinds of data.

Proposed Methodology: In the RAQ algorithm, all data are collected from the urban sensing system including air monitoring station data etc. The training dataset includes all the necessary features and is divided into subsets using bootstrap technology. A decision tree is constructed on each subset, and the classification is done by aggregating the results generated from all decision trees.

Year of Publication: 2014

Title: A Novel Air Quality Prediction Model Using Artificial Neural Networks.

Problem Statement: To develop the Air Quality prediction model for Sox, NOx and RSPM using meteorological parameters.

Proposed Methodology: ANN model was developed for Sox, NOx and RSPM prediction. The percentage of division of the dataset was as follows: Training Dataset- 75% Testing Dataset-15% Validation Dataset-15% Forward Selection Method is used for selecting the number and type of input variable in the model.

Year of publication: 2013

Title: Forecasting Criteria Air Pollutants Using Data Driven Approaches; An Indian Case Study

Problem Statement: To forecasting concentration of criteria pollutants such as Sox, NOx and RSPM one day in advance for Pune, India using ANN as well as GP and comparing them with respect to their accuracy of forecast.

Proposed Methodology: 3 models each of ANN & GP were developed i.e., ANN(Sox), ANN(NOx), ANN(RSPM), GP(NOx), GP(Sox), GP(RSPM). Out of the 7 previous days values, the most influential inputs were identified using correlation analysis and it was found that 3, 4, and 3 antecedent values prove to be the optimum inputs for Sox, NOx and RSPM models respectively.

Year of publication: 2012

Title: Short-Term Prediction of Air Pollution in Macau Using Support Vector Machines

Problem Statement: Using observed meteorological and pollutant data, SVM models for forecasting daily ambient air pollutant were constructed.

Proposed Methodology: In order to discern the relationship between meteorological data and pollutants, Pearson correlation coefficients was calculated. Parameters with coefficient value greater than 0.5 were selected as input in the SVM model. 5 different SVM models each for predicting the concentration of No2, So2, O3, SPM were developed based on the kernels used in them.

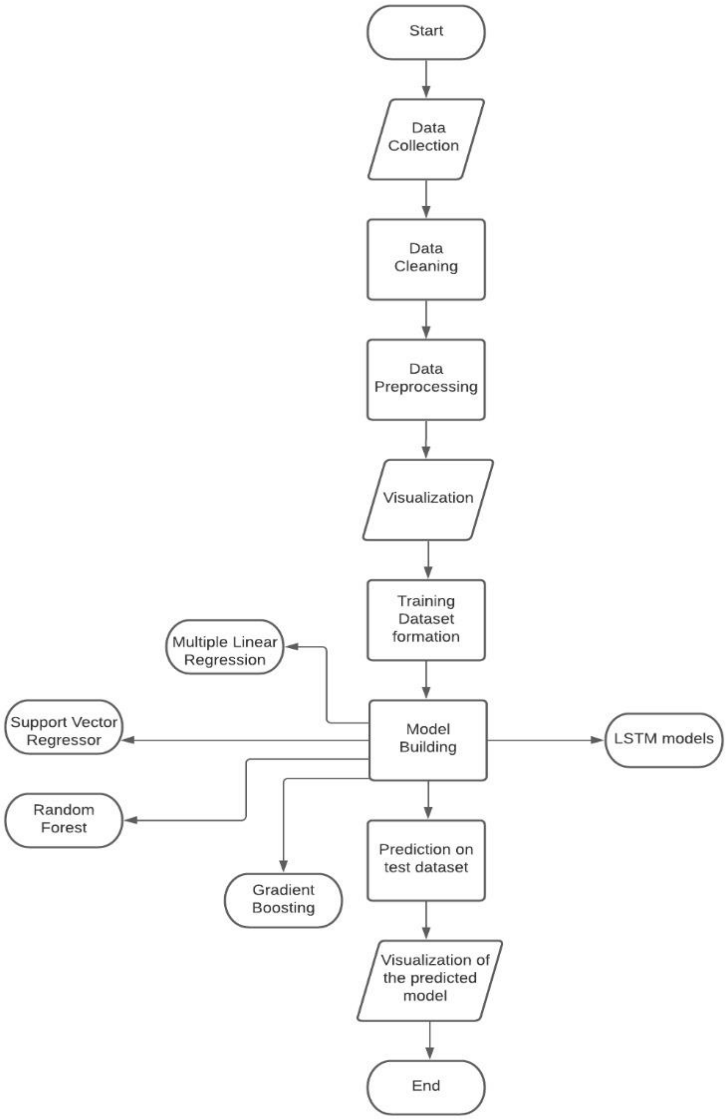
Year of publication: 2011

Title: Forecasting of air quality in Delhi using principal component regression technique

Problem Statement: To forecast the daily AQI value of Delhi, India based on previous day's AQI and meteorological parameters for four seasons - summer, monsoon, post monsoon, and winter.

*Proposed Methodology: P*revious day's AQI and meteorological parameters were used as independent features to forecast the AQI value using the Multiple Linear Regression model. Principal Component Regression was also used and given to the regression model (called PCR) which reduced the collinearity in the data and determined the relevant independent features.

4. Overall design of project



5. Features of project

To forecast the AQI values in Bangalore city using Long Short Term Memory neural network and various other machine learning models.

Languages Used

This project is made in Python language.
Following python libraries have been used to make the project:
Data Pre-processing: Numpy, Pandas
Visualization: Matplotlib, Seaborn
Machine Learning Models: Sklearn
LSTM Model: Pytorch

6. Proposed Methodology

1. Data Collection
2. Data Cleaning
3. Data Pre-processing
4. Visualization
5. Training Dataset Formation
6. Model Building
7. Prediction on testing dataset
8. Visualization of predicted model
9. Result

7. Algorithm Used

To forecast the AQI values, we built the following 4 Machine Learning models:

- Multiple Linear Regression
- Support Vector Regressor
- Random Forest
- Gradient Boosting

We also developed four versions of LSTM models and compared the result obtained from these models with the actual AQI values to find the best model.

8. Division of the work among students

Khushi Jain	Worked on making the LSTM Model 4, LSTM Model 3 and machine learning model Random Forest. Also performed Data pre-processing with the scrapped data and helped in making the project report.
Megha Agarwal	Contributed in making a LSTM Model 2 and machine learning model MLR and SVR. Also Performed web scraping with the selected dataset and worked in creating the project report.
Pavini Jain	Added on by making LSTM Model 1 and machine learning model Gradient Boosting. Also, performed data visualization with the trained dataset and helped in making the project report.

9. Results

Comparison between various models was done on the basis of R2 score, Mean absolute error and root mean squared error is shown in the below tables:

Comparison between various ML Models

MODEL	R2_SCORE	Mean absolute error	Root mean squared error
MULTIPLE LINEAR REGRESSION	0.81702036287260	0.15755965269436	0.19739750373716397
SUPPORT VECTOR REGRESSION	0.80630010071042	0.16675010409400	0.22212095937096654
RANDOM FOREST	0.75742992832956	0.19775706724581	0.24856698890550402
GRADIENT BOOSTING	0.78394965329306	0.17724614948669	0.23458612795481937

Comparison between various LSTM Models

MODEL	R2_SCORE	Mean absolute error	Root mean squared error
LSTM MODEL 1	0.766110725303	0.19053435	0.49583682
LSTM MODEL 2	0.773016977716	0.18732437	0.49213535
LSTM MODEL 3	0.771508592511	0.18938948	0.49295092
LSTM MODEL 4	0.802407929275	0.1743705	0.4753665

10. Conclusion

The Air quality index (AQI) or Air pollution index (API) is a standard method of informing the public about the severity of air pollution. Various researchers/environmental agencies have created a number of ways for determining AQI or API in the past, but there is no globally approved method that is adequate for all scenarios.

Because of the dynamic nature, volatility, and great unpredictability in location and time of pollutants and particles, predicting air quality is a difficult undertaking. Simultaneously, due to the recognised significant repercussions of air pollution on humans and the environment, the ability to model, predict, and monitor air quality is becoming increasingly vital, particularly in metropolitan areas.