Air Quality Prediction over Ahmedabad Using Machine Learning

Yash Dahima
Ahmedabad University, yash.d2@ahduni.edu.in

Abstract - **Air quality predictions can enable people to make informed decisions about their outdoor activities. The present work investigates 9 years of air pollution and meteorological data over Ahmedabad. An initial data exploration has been conducted to uncover the patterns in the dataset, and the parameters that have a direct impact on the air pollutant concentration have been identified. The data has been thoroughly processed and relevant features have been chosen based on correlation analysis and principal component analysis (PCA). Various machine learning models such as linear models, kernel-based models, ensemble models were developed and evaluated via metrics such as r^2 score, mean absolute error, mean absolute percentage error, root mean square error using scikit-learn library of python. Ensemble models were found to perform the best on the dataset.**

Index Terms – air quality, feature selection, exploratory data analysis, principal component analysis, linear models, kernel-based models, gradient boosting

INTRODUCTION

Air pollution is a major health issue worldwide, and India is one of the most polluted countries in the world. Also, Ahmedabad is among many highly polluted cities of the country. In today's world, it is crucial to monitor and forecast air quality, particularly in developing nations such as India. Machine learning-based prediction technologies are found to be the most effective means of studying these contemporary hazards, as opposed to conventional methods.

Human activities are responsible for the majority of air pollution, with sources such as industrial plants, automobiles, planes, burning of coal, kerosene and straw, and aerosol cans. Every day, hazardous pollutants like CO, CO_2, Particulate Matter (PM), NO_2, SO_2, O_3, NH_3, Pb, etc. are released into the environment. The airborne solid particles having size less than 2.5 μm are called $PM_{2.5}$ pollutants. The chemicals and particles that make up air pollution have a negative impact on the health of humans, animals, and plants. Air pollution can cause a range of severe illnesses in humans, including bronchitis, heart disease, pneumonia, and lung cancer. Poor air quality also contributes to other contemporary environmental issues such as global warming, acid rain, reduced visibility, smog, aerosol formation, climate change, and premature deaths.

LITERATURE SURVEY

The analysis of six years of air pollution data in Indian cities, with a focus on twelve air pollutants and the Air Quality Index (AQI) has been addressed in [1]. The dataset was preprocessed and cleaned, followed by the application of data visualization techniques to uncover hidden patterns and trends. Furthermore, the paper addresses data imbalance through resampling techniques and compares the performance of five popular ML models using standard metrics. The accuracy of various machine learning (ML) models was evaluated using classical statistical error metrics on both the train and test sets. The XGBoost model demonstrated the highest accuracy, while the SVM model had the lowest. To compare the performance of ML models, metrics such as MAE, RMSE, RMSLE, and R2 were evaluated. Based on these metrics, the XGBoost model was found to be the best overall performer, achieving optimal values during both training and testing phases.

In another work [2], the study presents machine learning models for analyzing air pollution using TAQMN data from Taiwan. The dataset spans from 2012 to 2017 and includes records from 76 air pollution stations. The focus of the study is on predicting particulate matter $PM_{2.5}$ levels using machine learning models, with evaluation based on statistical metrics such as MAE, MSE, RMSE, and R2. The results indicate that the proposed models outperform previous models, with actual and predicted values showing close agreement. Based on the evaluation, the study concludes that the gradient boosting regressor model is the most effective in forecasting air pollution on the TAQMN dataset.

DATASET GENERATION

The dataset is generated using the CAMS global reanalysis (EAC4) dataset. The original data was in a netCDF format which was transformed into an excel file using a python program. The dataset contains various meteorological as well as aerosol parameters over Ahmedabad from 2011 to 2019 at the interval of 3 hours. The parameters of the dataset are wind speed, temperature, humidity, precipitation, atmospheric pressure, amount of water vapor, boundary layer height and aerosol optical depth. The dataset was checked for missing values and no missing values were found.

EXPLORATORY DATA ANALYSIS (EDA)

The first few rows and columns of the dataset are shown in figure I. Different kinds of plots were generated to observe patterns in the data. Pair plots of all the parameters with the target variable were useful in observing the dependency of the target variable on different parameters.

	A	B	C	D	E	F	G	H	I
1	datetime	ws	temp	rh	dew_temp	precipitation	pressure	wv	blh
2	2011-01-01 05.30.00	2.74831796	11.0527344	61.081372	3.83377075	0.00014796	1007.85303	7.45163059	124.595856
3	2011-01-01 08.30.00	3.28579712	12.5096436	55.5465952	3.84979248	0	1009.77966	7.14379501	281.450348
4	2011-01-01 11.30.00	3.73198175	22.3202209	28.9550116	3.39859009	0	1010.1355	7.1469593	1059.9364
5	2011-01-01 14.30.00	3.8311882	25.0738831	20.7847345	1.07104492	0	1006.98138	7.51434898	1449.8009
6	2011-01-01 17.30.00	2.46180248	23.6787109	23.3570702	1.53250122	0	1006.1955	7.55980587	124.113068
7	2011-01-01 20.30.00	2.46064281	16.0066833	45.2044744	4.13839722	0	1007.55505	7.65532112	94.1943588
8	2011-01-01 23.30.00	2.82889056	13.9561157	51.6390683	4.1546936	0.00029593	1007.94019	7.68898201	129.797745
9	2011-01-02 02.30.00	3.01813817	12.3333435	55.9228979	3.78121948	0.00044389	1007.2713	7.5523262	126.817406
10	2011-01-02 05.30.00	2.74143744	11.4280396	57.1529657	3.24313354	0.00014796	1006.9549	6.97606754	118.642891
11	2011-01-02 08.30.00	3.06174135	13.6370239	50.4531263	3.52926636	0	1008.80823	6.59285355	223.289185
12	2011-01-02 11.30.00	2.88268256	22.9694824	30.3956238	4.65194702	0.00042616	1009.42767	6.65010548	894.594727

FIGURE I

FIRST FEW ROWS AND COLUMNS OF THE DATASET

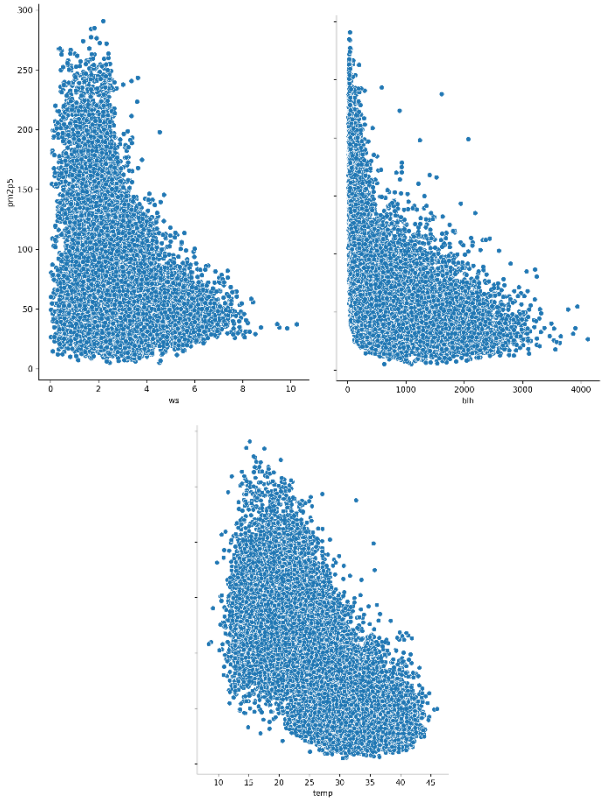

FIGURE II

SCATTER PLOTS OF $PM_{2.5}$ CONCENTRATION WITH (A) WIND SPEED (B) BOUNDARY LAYER HEIGHT (C) TEMPERATURE

As it can be seen in the figure II, lower $PM_{2.5}$ concentration is found at higher values of wind speed, boundary layer height and temperature. This is due to the fact that when the magnitude of these parameters is higher, there is more dispersion of pollutants. Hence, its concentration decreases.

FEATURE ANALYSIS AND SELECTION

Feature analysis and selection was performed using two techniques: correlation analysis and principal component analysis (PCA).

Correlation Analysis: Correlation coefficients were calculated for each pair of parameters and their values are shown in figure III.

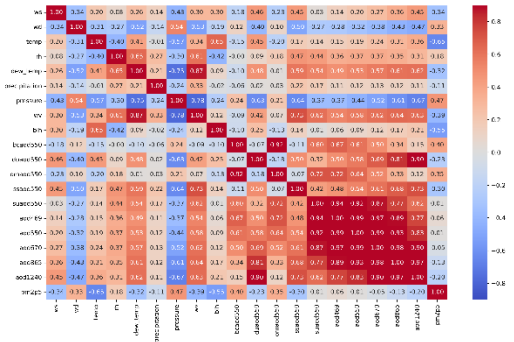

FIGURE III

CORRELATION MATRIX

Wind speed, temperature, boundary layer height, and pressure show a high level of correlation with the target variable, so they are important to build a model.

Principal Component Analysis (PCA): Each parameter was scaled by dividing it by its standard deviation to perform PCA.

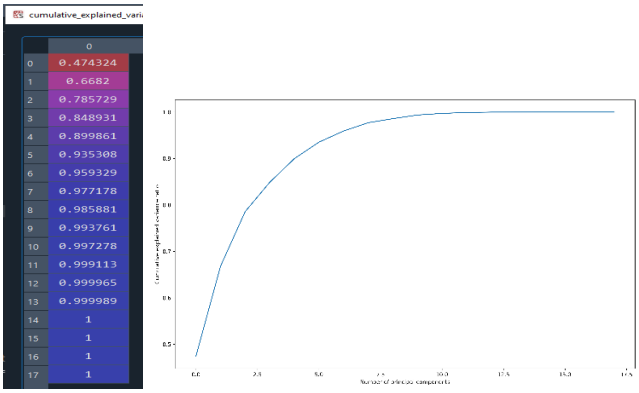

FIGURE IV

PCA

It is evident from figure IV that almost 90% of the variation in the data is explained by the first 5 principal components. Hence, the parameters contributing strongly to the first 5 principal components are to be selected to keep for developing a model. The features selected based on PCA and Factor Analysis for developing a model are: wind speed, temperature, relative humidity, pressure, water vapor, boundary layer height, and bcaod550.

MODEL DEVELOPMENT

Linear Models: Linear Regression (OLS - Ordinary Least Square), Ridge Regression, and Lasso Regression from scikit-learn python library have been used to develop models. Models Ridge Regression and Lasso Regression were developed with cross-validation by trying different regularizer strengths and data was fitted with the best value of regularizer.

Kernel-based Models: Ridge Regression model was developed using a periodic kernel with random search of the best parameters. 100 iterations were performed to choose the best combination of alpha, kernel length scale, and kernel periodicity. The model was fit with the dataset and evaluated using different metrics. Randomized Search provided the best values of alpha = 771.7, kernel length scale = 32.0, and kernel periodicity = 0.22 out of 100 iterations. The Gaussian Process Regressor model was developed using a periodic kernel to address daily and seasonal periodicity, and quadratic kernel to address trends in the dataset.

Ensemble Models: Four models extra-trees, random forest, and gradient boosting were fitted on the dataset using scikit-learn. They were run with the same random state and 10,000 number of estimators for the inter-comparison. Extra-trees performed best in this run. The models were again run with different parameters and gradient boosting performed best this time. There wasn't a significant difference in the performances of these models.

MODEL EVALUATION

Models were evaluated based on metrics such as r^2 score, mean absolute error (MAE), mean absolute percentage error (MAPE), root mean square error (RMSE) using scikit-learn library of python.

It is evident from table 1 that the linear model performances are not satisfactory as they are linear in nature and there is some periodicity present in the data. Hence, the models with periodic kernel functions were developed further.

The performances of models with periodic kernels was still not satisfactory as they were still not able to capture the pattern in the data accurately.

Finally the ensemble-based models showed good performance for the dataset. There wasn't a big difference in the performances of these models.

CONCLUSIONS

The task of air quality predictions over Ahmedabad using various machine learning models was attempted in this study. Dataset features for model development were selected based on correlation and PCA. The dataset was fitted using various classical machine learning algorithms such as linear models, kernel-based models, and ensemble models. The former two were not able to capture the patterns in the data accurately, whereas gradient boosting, random forest, and extra-trees regressors were able to do the same.

Models	RMSE	MAE	MAPE(%)	R^2
Linear (OLS)	28.33	21.69	36.03	0.56
Ridge	28.32	21.68	36.00	0.56
Lasso	28.34	21.70	36.06	0.55
Gaussian Process	40.50	32.40	44.50	0.1
Kernel Ridge	53.90	40.40	47.40	-0.2
Extra Trees	19.90	15.09	25.73	0.78
Random Forest	18.38	13.95	23.22	0.814
Gradient Boosting	18.20	13.63	22.19	0.818

TABLE I
MODEL EVALUATION METRICS

GITHUB Repo Link:
https://github.com/yash-dahima/CSE523-Machine-Learning-2023-Air-Quality-Prediction

REFERENCES

[1] Kumar, K., Pande, B.P. Air pollution prediction with machine learning: a case study of Indian cities. Int. J. Environ. Sci. Technol. (2022). https://doi.org/10.1007/s13762-022-04241-5

[2] Doreswamy HKS, Yogesh KM, Gad I (2020) Forecasting Air pollution particulate matter (PM2.5) using machine learning regression models. Procedia Comput Sci 171:2057–2066. https:// doi. org/ 10. 1016/j. procs. 2020. 04. 221

[3] Madhuri VM, Samyama GGH, Kamalapurkar S (2020) Air pollution prediction using machine learning supervised learning approach. Int J Sci Technol Res 9(4):118–123