Presentation on:

Application of Data Mining Techniques on Air Pollution of Dhaka City

Paper ID: 55

Proceedings of 2020 IEEE 10th International Conference on Intelligent Systems

Authors

- Al-Sadman Chowdhury
- Md. Shihab Uddin
- Md Rashad Tanjim
- Fariha Noor

Undergraduate Students, Department of ECE, North South University, Dhaka, Bangladesh

Dr. M. Rashedur Rahman

Professor, Electrical and Computer Engineering Department, North South University, Dhaka, Bangladesh

Table of Contents

Introduction

Air Pollution occurs when the level of air pollutants exceeds a certain limit. In our paper, we have used machine learning models to classify AQI level of different places of Dhaka city and we have used deep learning approaches using time series modeling to show in what way the air quality has decreased over the years.

- For the machine learning part, we have used decision tree, random forest, SVM, KStar, Ensemble selection, Multi-Layer Perception and bagging models.
- For the deep learning part, we only have used LSTM in two scenarios. One is for hourly prediction and the other is for daily prediction.

Background Study

It is important to know what has been done in the current field of work, to get an overall picture where the field currently stands.

- A. Kurt, B. Gulbagci, F. Karaca, and O. Alagha, “An online air pollution forecasting system using neural networks,” *Environment international*, vol. 34, pp. 592–8, 08 2008.

- P. Raj, “Prediction and optimization of air pollution-a review paper,” *International Journal for Research in Applied Science and Engineering Technology*, vol. 7, pp. 3896–3904, 05 2019.

Background Study (Cont.)

- G. Kaur, J. Gao, S. Chiao, S. Lu, and G. Xie, “Air quality prediction: Big data and machine learning approaches,” *International Journal of Environmental Science and Development*, vol. 9, pp. 8–16, 01 2018.

- T. Chiwewe and J. Ditsela, “Machine learning based estimation of ozone using spatio-temporal data from air quality monitoring stations,” 07 2016.

- R. Yu, Y. Yang, L. Yang, and G. Han, “Raq–a random forest approach for predicting air quality in urban sensing systems,” *Sensors*, vol. 16, p. 86, 01 2016.M. Delavar, A. Gholami, G. Shiran, Y. Rashidi, G. Nakhaeizadeh

Datasets

For our paper, we have used two datasets in our analysis.

- ***Weather Dataset***
  - We have collected the weather dataset for Dhaka city from 2016 to 2019 from the website https://en.tutiempo.net/.
  - It had a total of 1300+ instances and 19 attributes.

- ***Air pollution Dataset:***
  - The data set is collected from the site of National Oceanic and Atmospheric Administration (NOAA).
  - The data-set obtained has more than 31,000 instances.

Datasets (Cont.)

SL	Symbol	Interpretation
1	Y	Year
2	D	Day
3	T	Average Temperature (°C) *
4	TMax Temp	Max-Temperature (°C) *
5	TMin	Min-Temperature (°C) *
6	SLP	Atmospheric-Pressure-at-sea-level (hPa) *
7	H	Avg-Relative-Humidity (%) *
8	PP	Total Rainfall (mm) *
9	VV	Average Visibility (km) *
10	V	Average Wind Speed (km/h)
11	VM	Maximum Sustained Wind Speed (km/h) *
12	VG	Maximum Speed of Wind (km/h) *
13	RA	If there was there rain or drizzle
14	SN	If-there was snow in that month
15	TS	If there was any thunderstorm
16	FG	If there was any fog *
17	AQI Category	Air Quality Index **
18	NowCast	NowCast PM2.5 Concentration(ug/m3) *
19	Raw	Raw-PM2.5-Concentration (ug/m3) *

Weather Dataset

Standard AQI Level Implication

AQI	Air Pollution Level	Health Instructions	Cautionary Statement
0-50	Good	No health implications.	Normal Outdoor activity for everyone.
51-100	Moderate	Acceptable air quality. However, it can be harmful for hypersensitive people.	Caution for Hypersensitive people.
101-150	Unhealthy For sensitive groups	People with sensitive health condition can be affected to a large extent	Caution for children, elders and hypersensitive people
151-200	Unhealthy	Normal People may feel a bit uncomfortable while breathing while sensitive group can have a heart disease.	Sensitive people should avoid outdoor activities and general people should reduce outdoor activities
201-300	Very Unhealthy	Normal people will have a slight effect while sensitive people will be affected significantly	People should remain indoor unless it's an emergency.
300+	Hazardous	Healthy people can have a respiratory problem. Elders and the sick will be affected the most. Healthy people should also remain at home	Everyone should remain indoor and avoid physical exertion specially for the sensitive people

AQI Categorization and its Implication

Dataset Preprocessing

The raw dataset contained many missing as well as repeated values. We have preprocessed the data for correct analysis. The steps followed were:

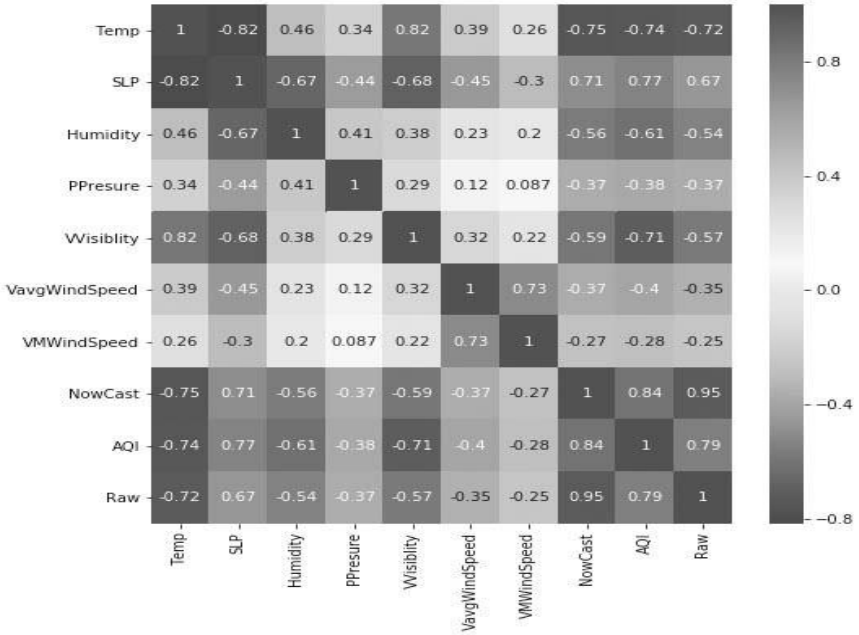
- **Conversion:**
  - we have obtained a total of more than 31000 instances.
  - It contained many continuous values.
  - we have converted those continuous values to discrete values, by applying a range.
- **Replacement:**
  - we have faced the issue of missing and negative values.
  - we replace the negative values with the monthly mean that we obtained directly with our dataset.

Dataset Preprocessing (Cont.)

- **Normalization of the Attributes:**
  - one attribute has the range of values in thousands, whereas another attribute has range of value between 0 and 1. Hence, the data was normalized.
- **Feature Selection:**
  - Two datasets have been merged into a single dataset.
  - We have calculated the gain ratio of different features and only selected those features, which have high gain ratio.
- **Correlation Matrix:**
  - We have filtered the dataset and then determined the correlation among different attributes.

Methodology

- *Correlation Matrix:*



Correlation Matrix

Methodology (Cont.)

- *Re-Sampling Data*

- $PM_{2.5}$ *NowCast Calculation*

$$NowCast = \frac{\sum_{i=1}^{12} w^{i-1} c_i}{\sum_{i=1}^{12} w^{i-1}}$$

$$I = \frac{C - C_{min}}{C_{max} - C_{min}} (I_{max} - I_{min}) + I_{min}$$

Where: I = AQI value,

C = pollutant concentration average for 24 hours,
$\boldsymbol{C_{min}}$ *= break-point for concentration that is* $\leq C$
$\boldsymbol{C_{max}}$ *= break-point for concentration that is* $\geq C$
$\boldsymbol{I_{min}}$ *= Correlating to* $\boldsymbol{C_{min}}$ *in terms of index-break point,*
$\boldsymbol{I_{max}}$ *= Correlating to* $\boldsymbol{C_{max}}$ *in terms of index-break point*

Methodology (Cont.) - Classifiers

- Decision Tree, Random Forest, SVM, Kstar, Ensemble method

Classifier	F-measure (Avg.)	ROC Area (Avg.)	Accuracy	Build Time (sec)
Decision Tree	0.913	0.965	91.48%	0.02
Random Forest	0.933	0.993	93.37%	0.86
SVM	0.779	0.837	77.43%	0.41
Kstar	0.883	0.976	88.47%	0.01
Bagging	0.900	0.989	90.52%	0.22
Ensemble Selection	0.890	0.982	89.42%	1.75
Multilayer Perceptron	0.854	0.956	85.79%	2.22

COMPARISONS OF DIFFERENT CLASSIFIERS WITH ACCURACY, F- MEASURE, ROC AREA AND MODEL BUILD TIME

Methodology (cont.)

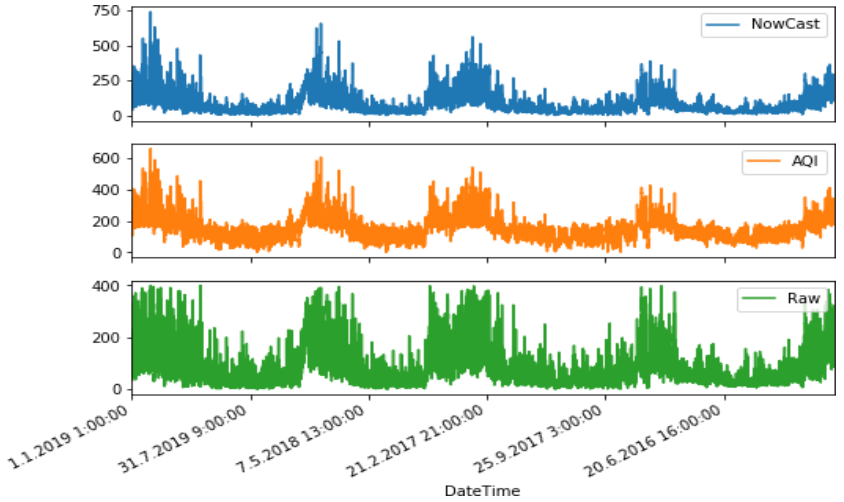
Among all classifiers the Random Forest gives the best result, so we have selected it for further analysis.

Hazardous	Very Unhealthy	Sensitive	Unhealthy	Moderate Good	Good
24	12	1	0	0	0
4	222	0	3	0	2
2	6	384	6	5	1
0	5	12	237	0	3
9	0	3	0	298	2
0	1	0	0	8	26

CONFUSION MATRIX OF RANDOM FOREST

Methodology (cont.) - Deep Learning

- LSTM for hourly prediction:
- LSTM for daily prediction



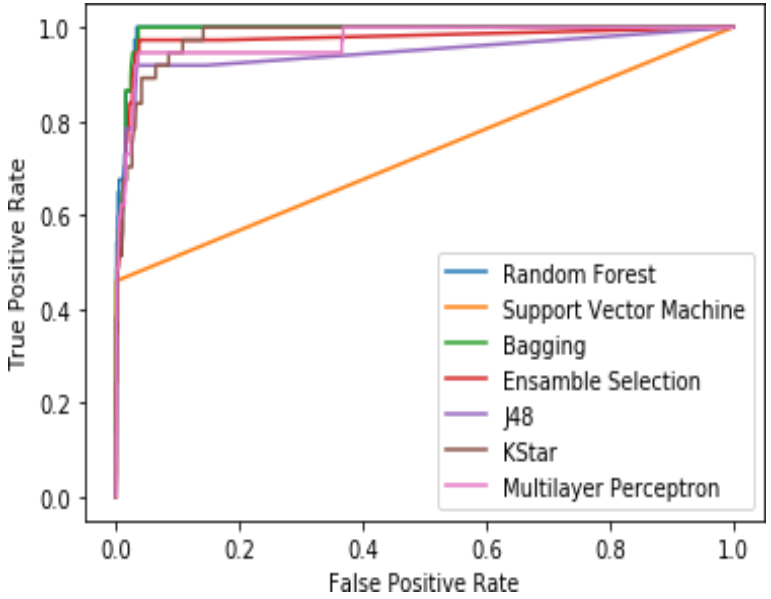
Time series analysis of AQI

Result Analysis – Classifiers

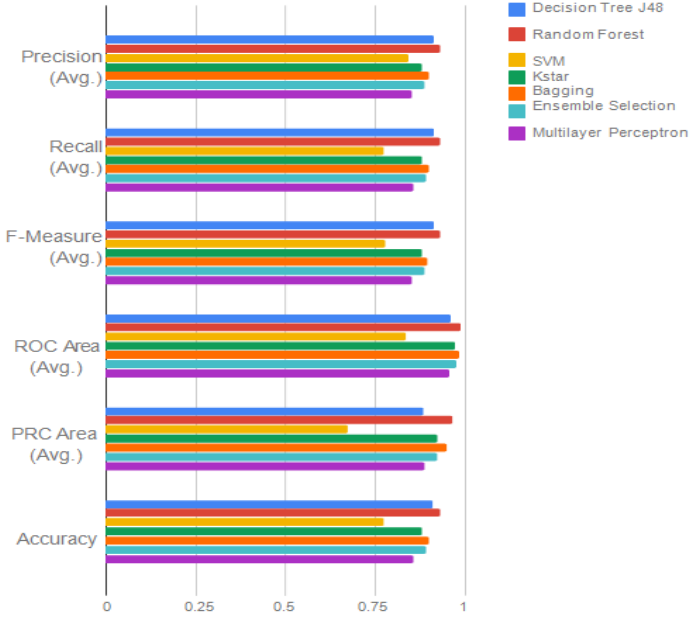
This table shows the different metrics that were considered for evaluating the models.

Rules	Class level
NowCast <= 62.47 and NowCast <= 34.37 and Raw > 133.87	Good
NowCast > 62.47 and NowCast > 140.8 and NowCast <= 248.7	Very Unhealthy
NowCast > 62.47 and NowCast <= 140.8 and Vvisiblity <= 2.9 and NowCast <= 82.4 and VMWindSpeed > 2.8	Sensitive
NowCast > 62.47 and NowCast <= 140.8 and Vvisiblity >2.9 and Tmin <= 26.1 and Raw <= 72.79 and Tmax <= 30.2	Unhealthy
NowCast > 62.47 and NowCast <= 140.8 and Vvisiblity > 2.9 and Tmin > 26.1 and SLP > 1004.7 and PPressure > 305.64	Moderate Good
NowCast > 62.47 and NowCast > 140.8 and NowCast > 248.73	Hazardous

Result Analysis (Cont.) – Classifiers



A receiver operating characteristics curve



Analysis of all the models used

Result Analysis (Cont.) – Classifiers

	TP Rate	FP Rate	Precision	Recall	F meas.	ROC	Class
	0.841	0.056	0.833	0.841	0.837	0.960	A
	0.889	0.048	0.856	0.889	0.872	0.977	B
	0.758	0.022	0.823	0.758	0.789	0.933	C
	0.842	0.028	0.818	0.842	0.830	0.958	D
	0.871	0.019	0.859	0.871	0.865	0.971	E
	0.807	0.023	0.854	0.807	0.830	0.953	F
Weight Avg,	0.841	0.037	0.841	0.841	0.841	0.961	

The receiver operating curve for the class level 'Hazardous' for all classifiers.

Here,

A = Very Unhealthy,

B = Hazardous,

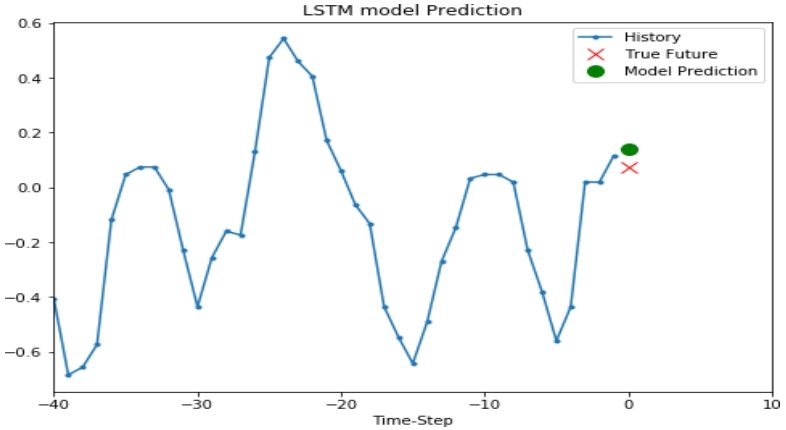
C = Moderate,

D = Good,

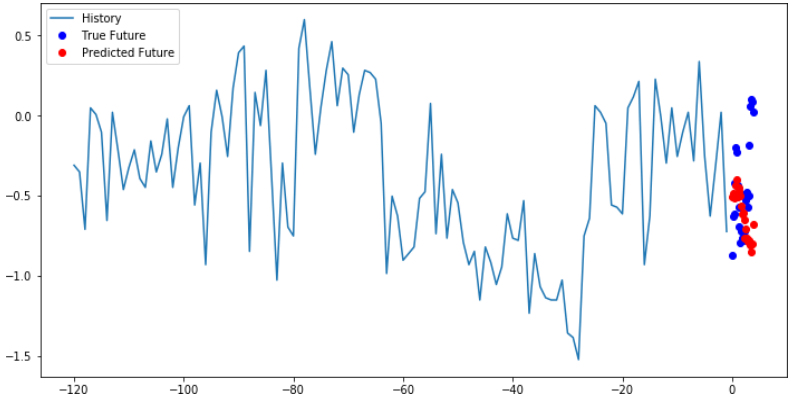
E = Moderate Unhealthy,

F = Unhealthy

Result Analysis (Cont.) - LSTM



Day wise air pollution prediction using LSTM



Hourly air pollution prediction using LSTM

Result Analysis (Cont.) – LSTM

- The table describes the mean square error of our LSTM models.

Model	MSE (Train)	MSE (Valid)
LSTM (Univariate Hourly Forecast)	0.059	0.026
LSTM (Multivariate Daily Forecast)	0.390	0.340

DIFFERENT ACCURACY METRICS

Conclusion

- We have applied different types of machine learning algorithms to predict the levels of pollutants in air based on previous air pollution and weather data.
- Random forest classifier gives the best accuracy of 93.37%.
- LSTM shows the pattern of increasing and decreasing of AQI based on different days and seasons.
- Analyzing this attributes we can more accurately predict the Air Quality Index and daily pollution rate.

Limitation of Study & Future Work

- More instance of data and attributes might gives a better analysis of our study on highly polluted city like Dhaka.
- In our future work, we want to use satellite images of different areas of Dhaka city, as well as include a few other major cities of Bangladesh to predict air pollution.
- We want to find patterns and relation how greenery and air pollution is related to each other.
- We will further explore our data-set to find interesting patterns such as the AQI level during the holidays.

References

Khan, S. Rahman, A. Haque, A. Chen, M. Hammond, S. Djordjevic´, and D. Butler, "Flood damage assessment for Dhaka city, bangladesh," *Flood Risk Management: Science, Policy and Practice: Closing the Gap*, p. 138, 11 2012.

"The World Bank." http://data.worldbank.org/indicator/SP.DYN.LE00. FE.IN. Accessed: 2019-12-11.

S. Hossain, "Rapid urban growth and poverty in Dhaka city," *Bangladesh e-Journal of Sociology*, vol. 5, 02 2008.

M. Rahman and A. Al-Muyeed, "Urban air pollution: a bangladesh perspective," 01 2005.

L. Miller and X. Xu, "Ambient pm2.5 human health effects—findings in china and research directions," *Atmosphere*, vol. 9, p. 424, 10 2018.

Y. Lin, J. Zou, W. Yang, and C. Q. Li, "A review of recent advances in research on pm2.5 in china," *International Journal of Environmental Research and Public Health*, vol. 15, p. 438, 03 2018.

P. K. Hopke, D. D. Cohen, B. A. Begum, S. K. Biswas, B. Ni, G. G. Pandit, M. Santoso, Y.-S. Chung, P. Davy, A. Markwitz, *et al.*, "Urban air quality in the Asian region," *Science of the Total Environment*, vol. 404, no. 1, pp. 103–112, 2008.

A. Kurt, B. Gulbagci, F. Karaca, and O. Alagha, "An online air pollution forecasting system using neural networks," *Environment international*, vol. 34, pp. 592–8, 08 2008.

P. Raj, "Prediction and optimization of air pollution-a review paper," *International Journal for Research in Applied Science and Engineering Technology*, vol. 7, pp. 3896–3904, 05 2019.

G. Kaur, J. Gao, S. Chiao, S. Lu, and G. Xie, "Air quality prediction: Big data and machine learning approaches," *International Journal of Environmental Science and Development*, vol. 9, pp. 8–16, 01 2018.

T. Chiwewe and J. Ditsela, "Machine learning based estimation of ozone using spatio-temporal data from air quality monitoring stations," 07 2016.

R. Yu, Y. Yang, L. Yang, and G. Han, "Raq–a random forest approach for predicting air quality in urban sensing systems," *Sensors*, vol. 16, p. 86, 01 2016.M. Delavar, A. Gholami, G. Shiran, Y. Rashidi, G. Nakhaeizadeh,

Dhaka Weather Datatset, 2016 to 2019, Tutiempo Network, S.L., 2019. [Online]. Available: https://en.tutiempo.net/climate/ws-419230.html. Accessed: 2019-12-11

Air Pollution Datatset, 2016 to 2019, National Oceanic and Atmospheric Administration., USA. [Online]. Available: https://www.airnow.gov/index.cfm?action=airnow.global_summary#Bangladesh$Dhaka. Accessed: 2019-12

R. Rashu, S. T. Jishan, N. Haq, and M. Rahman, "Implementation of optimum binning, ensemble learning and re-sampling techniques to predict student's performance," *International Journal of Knowledge Engineering and Soft Data Paradigms*, vol. 5, p. 1, 01 2015.

Rashad Tanjim , Application of Data Mining Techniques on Air Pollution of Dhaka City, (2020), GitHub repository, https://github.com/RashadTanjim/Application_of_Data_Mining_Techniques_on_Air_Pollution_of_Dhaka_City.git

Thank You!

Any Question?