A novel feature engineering algorithm for air quality datasets

Raja Sher Afgun Usmani[1], Wan Nurul Farah binti Wan Azmi[2], Akibu Mahmoud Abdullahi[3], Ibrahim Abaker Targio Hashem[4], Thulasyammal Ramiah Pillai[5]
[1]School of Computing & IT, Taylor's University, Malaysia
[2]School of Biosciences, Taylor's University, Malaysia
[3,4,5]Centre for Data Science & Analytics, Taylor's University, Malaysia

Article Info

Article history:

Received Jan 6, 2020
Revised Mar 15, 2020
Accepted Mar 26, 2020

Keywords:

Air pollution
Air quality
Air quality monitoring station
Data cleaning
Feature engineering

ABSTRACT

Feature engineering (FE) is one of the most important steps in data science research. FE provides useful features to be used later in the study. Due to climate change, the research focus is moving towards air quality estimation and the impacts of air pollution on health in Malaysia. Malaysia has 66 air quality monitoring (AQM) stations, and the air quality data for research is provided in an excel worksheet format by the Department of Environment, Malaysia. The data generated by the AQM stations is in a raw custom format, and it is virtually impossible to clean and engineer this data manually due to the sheer number of files. Hence, we propose a novel feature engineering algorithm to transform and combine this data into a useable format. The results show that the proposed feature engineering algorithm was able to efficiently extract and combine the hourly and daily values for pollutant and meteorological variables in useful row format. This algorithm will help all the researchers using the data from the AQM station in Malaysia as well as other countries using the same AQM station. The implementation of the feature engineering algorithm is also available to use at GitHub (https://github.com/rajasherafgun/featureengineeringaq) under AFL-3.0 license.

Corresponding Author:

Raja Sher Afgun Usmani,
School of Computing & IT,
1 Jalan Taylor's, Taylor's University,
Subang Jaya, Malaysia.
Email: rajasherafgunusmani@sd.taylors.edu.my

1. INTRODUCTION

Feature engineering is one of the most critical tasks performed by data scientists and researchers in general. Feature engineering provides good features, giving the flexibility to choose between various algorithms. The flexibility of choice means less complex algorithms will provide good accuracy with good features [1]. Hence, on average, data scientists spend 80% of their time collecting, capturing, cleaning, and organizing data [2]. Moreover, 75% of the data scientists report that cleaning and engineering data is the least enjoyable part of their work [2].

Feature engineering process can be applied to the data at any stage. It can transform raw data into useable data, as well as already engineered data into useable data for a specific task. The process of feature engineering consists of the transformation of collected/recorded parameters, generation of new parameter values from available features or patterns, extraction of features from raw data, selection of features based on a specific criteria, analysis and evaluation of the usefulness of features and automated methodologies for generating and selecting features [3]. Another definition of feature engineering is to use a data mining algorithm to extract features from raw data using domain knowledge [4]; hence, classifying feature engineering to be a domain-specific process. Feature engineering is applied to virtually every field, including image processing [5], signal processing [6], natural language processing [7], estimation, and prediction, among others.

Recently, climate change has forced researchers to study air pollution and other major causes of climate change. The air quality data is collected using various devices and air quality monitoring (AQM) stations. The air quality monitoring in Malaysia was carried out through a private company known as Alam Sekitar Malaysia Sdn Bhd (ASMA) until recently. ASMA was appointed by the Department of Environment (DOE), Malaysia and the Malaysian Meteorological Department (METMalaysia). ASMA was responsible for collecting, processing, analyzing, and distributing the air pollutant measurements. There are 66 AQM stations across Malaysia, 14 Manual **S**ampling (High Volume Sampler) stations were operated by METMalaysia, and ASMA [8] operated 52 continuous air-quality monitoring (CAQM) stations. The DOE has increased the number of CAQM station to 68 [9]. The location-wise details of the CAQM stations are provided in Figure 1. Under the new Environmental Quality Monitoring Programme (EQMP), the DOE is closely monitoring the environmental parameters. The new system provided by private contractor Transwater gathers and stores real-time data on river, sea, and air conditions.

Figure 1. Air quality monitoring stations in Malaysia

Air pollution is the driving force behind climate change, and it is considered to be one of the biggest environmental challenges faced by humanity in the 21st century [10]. The primary motivation behind air pollution studies is the health impacts involving air pollution. The air pollutants have a severe and harmful effect on health, and it has become a serious global threat to human health and welfare [11]. In Malaysia, the rapid economic growth and the commitment to achieve the industrial country status under Vision/Wawasan 2020 has contributed to the degradation of the urban environment and air quality. Industrialization and haze episodes combined, the air pollution has turned from bad to worse in recent years [12]. The air pollution in Malaysia is affecting ecosystems, forest species, agricultural crops [13], and human health [14].

Apart from the motivations to study air pollution, the motivation to propose this feature engineering algorithm is the sheer amount of information in data generated by AQM stations. The AQM station data is recorded and exported in excel worksheet format every month. The data is presented in a custom format by the AQM station reporting system, therefore, it requires a custom solution or manual work to clean and process. For every year, 12 excel worksheet files are generated with 23 sheets in each excel worksheet. With 66 CAQM stations in Malaysia, it makes 18,216 sheets in 792 excel worksheets for one year only. This feature engineering work is part of a research project which is using 11 years of data from CAQM stations, with 156,170 sheets in 6,790 worksheet files. This makes it virtually impossible to do feature engineering manually as it will take a considerable amount of time and effort to get this data in a simple row format. Due to these hurdles and challenges, we propose a novel feature engineering algorithm to engineer the data provided by AQM stations to a usable form. Recently, the data generated by AQM stations is being used in various research studies to forecast air pollution, study seasonal variation and spatial distribution of air pollution, the trend analysis and mitigation of air pollution, and the economic and health impacts of air pollution. The details of these studies are provided in Table 1.

The main contribution of this paper is to provide a novel feature engineering algorithm for researchers working in the domain of air pollution. The proposed algorithm takes all the files as input and provides usable CSV files efficiently. This algorithm will help all the researchers using the data from the AQM station in

Malaysia as well as other countries using the same AQM station. The rest of the paper is organized as follows: In Section 2, presents the data used as input for our study and the feature engineering algorithm. Section 3 includes the results, discussion and future work, and lastly, conclusions are presented in Section 4.

Table 1. Studies using DOE air quality data 2010-2020

Study Year	Aim/Objective	Location	Dataset Years
2020 [15]	Forecasting of Air Pollution Index (API)	Cheras, Kuala Lumpur	2012-2017
2019 [16]	Identification of main challenges in ozone mitigation	Petaling Jaya, Shah Alam and Cheras	2012 to 2014
2018 [17]	Evaluating the variability of air pollutants and meteorological conditions at Langkawi Island	Langkawi	1999-2011
2018 [18]	Investigating the level of PM10 due to seasonal variation in Sabah	Sabah	2012
2018 [19]	exploring the concentration of air pollutants in monsoon season	Kemaman, Kertih and Kuala Terengganu	1999-2011
2017 [20]	Prediction of spatial variation of Air Pollution in Johor	Pasir Gudang, Johor	-
2017 [21]	Description of Air Pollution concentrations in Klang	Klang City	2012-2014
2017 [22]	Spatial distribution of concentration of Ozone	Shah Alam, Kajang, Petaling Jaya and Port Klang	2014
2016 [23]	Investigates the variability of PM2.5 in industrial	Klang Valley	Aug. 2011-Jul
2015 [24]	Detection of daily PM10 concentration	Klang, Kuala Selangor and Petaling Jaya	2005-2010
2015 [25]	Investigates the trend of Air Pollution in Klang Valley	Klang Valley	
2015 [26]	Detect changes of Air Pollution and its impact of on human health	Selangor	2013
2014 [27]	Explores the effects of haze days on daily mortality in Klang Valley	Klang Valley	2000-2007
2014 [28]	Predicts the concentrations of Air Pollution in Malaysia	Pasir Gudang, Kuching, Bukit Rambai, Tasek, Nilai, Klang, Balok Baru, Pengkalan Chepa, Paka, and Labuan	2005–2011
2014 [29]	Predicts Air Quality Index using Artificial Neural Network and Multiple Linear Regression	Pasir Gudang, Bukit Rambai, Nilai, Johor Bahru, Bachang, Muar, Seremban, and Tampoi	2005–2007
2014 [30]	Assessed the economic value due to haze related illiness	Kuala Lumpur	2005, 2006, 2008, 2009
2013 [31]	Describes spatial and temporal variation of Air Pollutants	Petaling Jaya,Malacca city centre and Kuching	2000–2010
2013 [32]	Evaluates the variabiity of ozone level in Malaysia	Seberang Perai, Pulau Pinang Jerantut, Pahang Bakar Arang, Kedah Kajang, Selangor	2009
2012 [33]	Influence of meteorological conditions on a daily PM10 and NO2 pollutants	Shah Alam, Johor Bahru, and Kuching	2007-2009
2012 [34]	Description of Air Pollution changes in Klang Valley	Klang Valley	2001-2009
2012 [35]	Explores the costs and benefits of Foreign Direct Investment (FDI) in the Malaysian	-	-
2012 [36]	Concentration and variability of ozone in Klang Valley	Klang Valley	2004-2008
2012 [37]	Examines the spatial patterns and sources of Air Pollutants in Malaysia	Kuching, Sibu, Kota Kinabalu and Tawau, Klang, Shah Alam, Ipoh and Johor Baharu	2008-2009
2011 [38]	Relationship between the changes of PM10 in monsoon season in Klang Valley	Klang Valley	2003-2006
2011 [39]	The trend of Air Quality	Kajang, Nilai and Banting	1996-2006
2010 [40]	The trend of Air Quality in Klang Valley	Klang Valley	1997-2006

2. METHOD

2.1. Data

DOE, Malaysia provides the data used for input in this algorithm. The data is provided in a specific directory structure, as displayed in Figure 2. This feature engineering work is part of a research project which is using 11 years of data from CAQM stations, with 156,170 sheets in 6,790 worksheet files. Inside the files, the data is stored in a custom format. There are 23 sheets in each file. The first sheet contains a monthly summary of air contaminants. The other 22 sheets contain hourly readings and daily average for each air pollutant and meteorological variable. Table 2 provides the detail of these sheets.

The generated file names contain information about the year, month, and AQM station, as shown in Figure 2. This filename structure is advantageous because it helped to extract the information by year, month, and AQM station.

Table 2. Sheet & variable details

Sheet#	Variable Name	Sheet#	Variable Name	Sheet#	Variable Name
1	Monthly Report Summary	9	NO_2	17	SO_2 API
2	CO	10	Methane (CH4)	18	Total API
3	CO 8 Hour Running average	11	NmHC	19	Wind Direction & Wind Speed xxM (uofM) Avg
4	Ozone	12	Total Hydrocarbons	20	Ultraviolet-B
5	PM_{10}	13	CO API	21	Wind Direction & Wind Speed
6	SO_2	14	NO_2 API	22	Humidity (%) Avg
7	NOx	15	Ozone API	23	Wind Speed 10m Avg
8	NO	16	PM_{10} API		

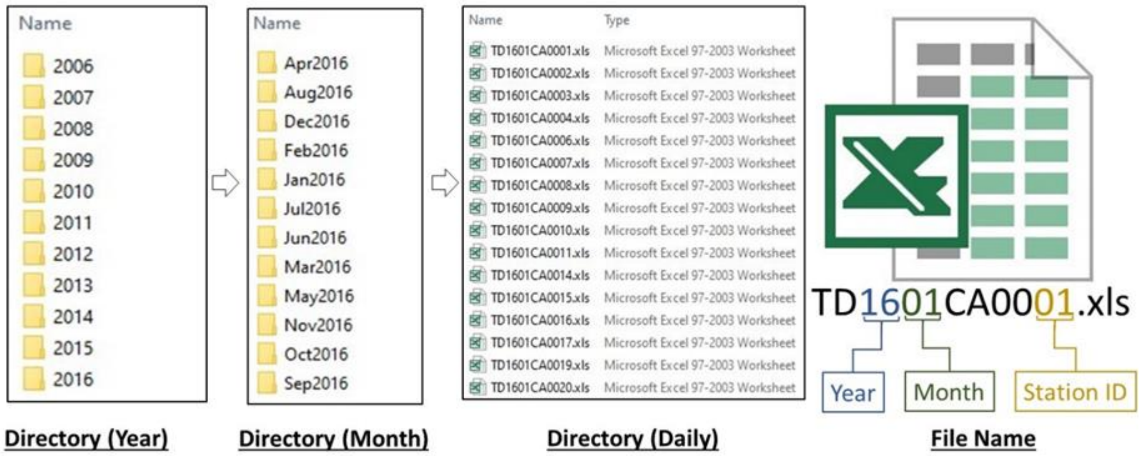

Figure 2. Directory & file name structure of dataset

2.2. The algorithm

The feature engineering algorithm is designed to be generic as it has to handle multiple pollutants. The algorithm uses two data structures, *AirPollutionData* and *AirPollutionDataCombined*. *AirPollutionData* is used to collect hourly information about a variable, station wise and *AirPollutionCombinedData* collects the daily average of all variables, station wise. Station ID is a unique number that represents a station. The data structures are shown in Figure 3.

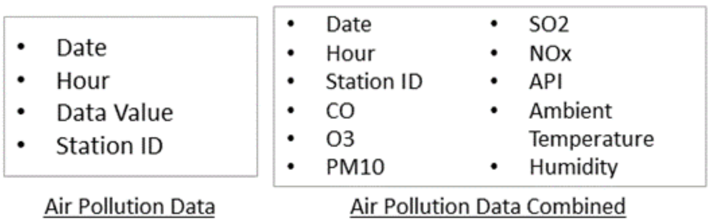

Figure 3. Data structures used in algorithm

The algorithm is divided into two parts, i.e., *LoadData* and *LoadDataFromExcel*. LoadData is the entry point of the algorithm, it takes all the file paths as input, and the output is engineered Comma Separated Values (CSV) files. LoadData is shown in Algorithm 1. The LoadData algorithm goes through all files, validates, extracts the information, populates appropriate lists and creates CSV files. It creates monthly CSV files for each station by combining the data in pollutant lists by date, hour, and station ID. It also creates a combined CSV file for all stations by combining the data in pollutant lists by date and Station ID.

```
Algorithm 1: Load Data
Data: All file names
Result: CSV files for Hourly and Daily air quality data
Function LoadData
  allFiles = Load all file names
  pollutantNames = Load all pollutant names
  foreach fileName in allFiles do
    if fileName[0] != 'T' OR fileName[1] != 'D'
      then
        continue;
    end
    year = Get year from fileName;
    month = Get month from fileName;
    stationID = Get stationID from fileName;
    foreach pollutantName in pollutantNames do
      LoadDataFromExcel(fileName,
      pollutantName, listToSaveIn, stationID,
      year, month);
    end
    CombinedList = Combine all pollutant lists,
    join by date, hour and stationID
    Create CSV for CombinedList, join by date
  end
  Create CSV for DailyListCombined, join by date,
  stationID
end
```

```
Algorithm 2: Load Data from Excel
Data: fileName, pollutantName, listToSaveIn,
      stationID, year, month
Result: Populate Hourly and Daily Lists
Function LoadDataFromExcel(fileName,
  pollutantName, listToSaveIn, stationID, year, month)
  DataRows = Load Data From File using fileName
  for i=1;i<=24;i++ do
    apObject=Read Hourly Values;
  end
  Add apObject to listToSaveIn
  if apObject exists in DailylistCombined then
    Populate pollutant information
  else
    Populate apObject with date, stationId and
    pollutant information
    Add to DailylistCombined
  end
end
```

3. RESULTS & DISCUSSION

Air pollution is one of the significant causes of climate change. Air pollution poses a variety of challenges for humans, ranging from economic to health. Researchers nowadays are focusing on the research on air quality and it's health impact. It is helpful that the air quality data is readily available in most countries. In Malaysia, this data is generated by AQM station in the form of excel worksheets. These excel worksheets contain hourly readings for all the recorded pollutants and meteorological variables. As this data is provided in a specific format, it needs to be engineered before it can be used in any research. As discussed in section 2, the sheer amount of information in these files need considerable time and resources to clean and engineer. The second part of the feature engineering algorithm is the part where we read the data and create two types of lists. The first list has hourly values for the pollutant, and the second part has the daily averages of the pollutant. It read the hourly data and saves in the respective list and then checks whether this data is already in the DailyListCombined, if it exists, it will update the entry with the pollutant value. If it does not exist, it will create a new entry in the DailyCombined List.

The algorithm will take the thousands of files from the dataset with custom format and extract, clean and combine the data into separate monthly useable CSV files with hourly values for the whole month. It also creates a combined file with all the data from all stations, all years and all months with daily averages. The example data is shown in Figure 4 and the example output is shown in Figure 5.

Carbon Monoxide (ppm) Avg
Hourly Summary

Date	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	Daily Avg
1	.50	.54	.49	.47	.42	.89	1.22	1.32	.77	.44	.42	.34	.28	CAL	.30	.31	.34	.44	.51	.53	.54	.53	.71	.62	.56
2	.60	.53	.64	.73	.46	.80	.70	.82	.66	.38	.31	.29	CAL	.25	.22	.23	.25	.28	.31	.41	.54	.58	.56	.45	.48
3	.43	.38	.22	.19	.16	.18	.27	.70	.30	.14	.17	CAL	.21	.17	.25	.20	.30	.47	.37	.36	.42	.49	.50	.38	.32
4	.39	.43	.49	.49	.42	.55	.62	.91	.48	.33	CAL	.28	.30	.27	.29	.30	.28	.31	.34	.35	.38	.42	.39	.38	.41
5	.32	.25	.27	.28	.26	.31	.57	.79	.47	CAL	.32	.31	.32	.29	.26	.25	.26	.30	.37	.33	.39	.41	.37	.37	.35
6	.36	.35	.32	.34	.34	.28	.43	.83	CAL	.32	.27	.26	.24	.24	.22	.24	.23	.28	.31	.31	.37	.35	.37	.34	.33
7	.48	.34	.38	.44	.38	.49	.79	CAL	.84	.29	.23	.22	.22	.22	.22	.29	.27	.31	.33	.26	.32	.36	.37	.38	.37
8	.38	.28	.32	.29	.33	.53	CAL	1.01	.75	.34	.25	.24	.26	.27	.31	.27	.25	.29	.32	.36	.35	.37	.45	.56	.38
9	.42	.30	.30	.39	.32	CAL	.50	.93	.59	.30	.28	.31	.33	.40	.39	.43	.45	.54	.59	.63	.79	.90	1.38	1.33	.56
10	.58	.60	.66	.61	CAL	.54	.55	.88	.58	.34	.26	.28	.26	.20	.24	.10	.13	.18	.25	.20	.17	.16	.21	.27	.36
11	.31	.43	.32	CAL	.28	.32	.61	1.22	.82	.37	.33	.35	.36	.41	.38	.33	.32	.39	.48	.59	.77	1.05	.73	.72	.52
12	.79	.54	CAL	.64	.61	.54	.72	1.92	1.90	.58	.40	.36	.36	.31	.28	.27	.31	.30	CAL	CAL	.60	.65	.80	.78	.65
13	.85	CAL	.75	.48	.42	.47	.77	2.44	.84	.47	.34	.32	.29	.27	.24	.26	.35	.44	.54	.41	.71	.92	1.09	1.27	.65
14	CAL	1.06	1.04	1.14	1.09	1.12	1.14	2.51	1.90	.66	1.34	.17	.13	.17	.19	.20	.21	.19	.35	.40	.34	.25	.45	CAL	.73
15	.63	.68	.26	.23	.27	.28	.27	.44	.46	.47	.34	.29	29	.30	.27	.25	.25	.33	.38	.43	.47	.64	CAL	.67	.39
16	.61	.55	.41	.29	.75	.77	.83	1.05	.72	.32	.25	.25	.21	.21	.23	.20	.27	.41	.55	.78	1.22	CAL	1.56	.98	.58

MonthlyReportSummary | CO Hourly | CO RAvg Hourly | O3 Hourly | PM10 Hourly | SO2 Hour ...

Figure 4. Example CO monitoring data in excel workbook

StationID	Year	Month	Day	CO	O3	PM10	NOx	NO2	NO	SO2	Total API	Ambient Temp	Humidity
1	2006	1	1	0.99087	0.007522	36.72727	0.033217	0.010783	0.022435	0.001682	37.125	25.625	86.91666
2	2006	1	1	0.263478	0.012826	24.08696	0.002783	0.002348	0.000435	0.000318	23.625	26.09583	91.29166
3	2006	1	1	0.92087	0.018714	31.08333	0.015087	0.010696	0.004391	0.000143	41	26.52083	86.375
4	2006	1	1	0.121429	0.006565	33.25	0.003435	0.001739	0.001696	0.000636	37.08333	24.80417	87.04166
5	2006	1	1	1.115652	0.022087	30.73913	0.035043	0.017261	0.017783	0.002696	44.375	25.92917	87.6
6	2006	1	1	0.470435	0.018565	51.58333	0.02287	0.01213	0.010739	0.001478	60.16667	24.31667	85.5
7	2006	1	1	0.348261	0.005174	34.25	0.003043	0.00213	0.000913	0	38.95833	24.525	84.54166
8	2006	1	1	0.518261	0.012957	30.875	0.017826	0.011348	0.006478	0.00087	36.25	26.48333	81.16666
9	2006	1	1	0.822609	0.009391	38.25	0.033652	0.014652	0.018957	0.003909	49.04167	26.37083	81.29166
10	2006	1	1	0.46087	0.015913	38.875	0.012391	0.007696	0.004696	0.001913	46.58333	25.15833	85.125
11	2006	1	1	0.734546	0.011727	45.45454	0.034636	0.022045	0.012773	0.002667	53.20833	26.9	82.21739
14	2006	1	1	0.276957	0.010913	27.29167	0.003739	0.002957	0.000783	0.001435	31.5	25.80417	83.58334

Figure 5. Example output data

We propose a novel feature engineering algorithm to save time and resources. To give an example of the amount of feature engineering required to process air quality datasets, this study is a part of research project to find the health impact of air pollution, using the data from AQM stations. To extract the Air Pollution Index and major 6 pollutants i.e. PM_{10}, NO_2, NO, O_3, CO and SO_2. We have to go through 48,048 sheets to get hourly and daily averages. This will require a considerable amount of time and resources to engineer them into a usable form. Figure 5 shows the sample output of our algorithm. We also use CSV as our output format as it is considered to be a standard format, and no propriety software is needed to read, parse, and use a CSV. Also, the best advantage of CSV is that it is smaller in size and processed by almost all existing applications.

The CAQM stations used in Malaysia are a product of MET One Instruments, Inc. CAQM stations by MET One Instruments are used all around the world, and ASMA is the primary distributor of MET One Instruments in Malaysia [41]. Hence, the algorithm provided in this study is useable for air pollution researchers around the world. In this study, we found 27 studies using AQM stations data in the last 11 years only. Our feature engineering algorithm will help many researchers to save time and resources. The DOE has also increased the number of AQM stations. Hence, more data is available in a variety of locations in Malaysia for research.

This algorithm is designed to extract information about major air pollutants (PM_{10}, O_3, CO, NO_x, NO_2, NO, SO_2 and API) and meteorological variables (Ambient Temperature, Humidity), but it can easily be expanded to extract all the air pollutants and meteorological variables from all 22 sheets.

4. CONCLUSION

Air pollution is one of the biggest environmental challenges in the 21st century, and it has a substantial economic and health effect. As with the rest of the world, air pollution is becoming a significant problem in Malaysia due to industrialization, a huge trend of using private cars, and haze episodes. Due to these reasons, air pollution has become a significant research area for researchers in Malaysia. These researchers are using the air quality datasets provided by the DOE and powered by MET One Instruments, Inc. The sheer number of files and information in these air quality datasets are overwhelming. In this study, we propose a feature engineering algorithm to extract, clean, and combine these features efficiently. The provided algorithm will save time and resources for researchers using air quality datasets in Malaysia as well around the world. The implementation of the feature engineering algorithm is also available to use at GitHub (https://github.com/rajasherafgun/featureengineeringaq) under AFL-3.0 license.

ACKNOWLEDGEMENTS

This research is funded by Taylor’s University under the research grant application ID (TUFR/2017/004/04) entitled as “Modeling and Visualization of Air-Pollution and its Impacts on Health”. We are also thankful to Department of Environment, Malaysia for providing the air quality monitoring station datasets.

REFERENCES

[1] J. Thanaki, *Python natural language processing*. Packt Publishing Ltd, 2017.
[2] S. Ozdemir and D. Susarla, *Feature Engineering Made Easy: Identify unique features from your dataset in order to build powerful machine learning systems*. Packt Publishing Ltd, 2018.
[3] G. Dong and H. Liu, *Feature engineering for machine learning and data analytics*. CRC Press, 2018.

[4] Feature Engineering "Feature Engineering," *Feature Engineering - Wikipedia*, 2020. [Online]. Available: https://en.wikipedia.org/wiki/Feature_engineering. [Accessed on Mar 26, 2020].

[5] A. F. H. Alharan, H. K. Fatlawi, and N. S. Ali, "A cluster-based feature selection method for image texture classification," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 14, no. 3, pp. 1433–1442, 2019.

[6] M. T. S. Al-Kaltakchi, H. A. A. Taha, M. A. Shehab, M. A. M. Abdullah, "Comparison of feature extraction and normalization methods for speaker recognition using grid-audiovisual database," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 18, no. 2, pp. 586–598, 2020.

[7] A. Adeleke, N. A. Samsudin, Z. A. Othman, and S. K. A. Khalid, "A two-step feature selection method for quranic text classification," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 16, no. 2, pp. 730–736, 2019.

[8] "Sources of air pollution in Malaysia - Jabatan Alam Sekitar," Department of Environment, Malaysia, 2013. [Online]. Available: http://www.doe.gov.my/portalv1/wp-content/uploads/2013/06/General-Information-of-Air-Pollutant-Index.pdf. [Accessed on Mar 26, 2020].

[9] "Air Pollutant Index of Malaysia," Department of Environment, Malaysia, 2020. [Online]. Available: http://apims.doe.gov.my/public_v2/home.html. [Accessed on Mar 26, 2020].

[10] P. Matson, "Environmental challenges for the twenty-first century: Interacting challenges and integrative solutions," *Ecol. Law Q.*, vol. 27, no. 4, pp. 1179–1190, 2001.

[11] A. B. Hansen, *et al.*, "Long-term exposure to fine particulate matter and incidence of diabetes in the Danish Nurse Cohort," *Environ. Int.*, vol. 91, pp. 243–250, 2016.

[12] M. T. Latif, *et al.*, "Impact of regional haze towards air quality in Malaysia: A review," *Atmos. Environ.*, vol. 177, pp. 28–44, 2018.

[13] S. Ishii, F. M. Marshall, J. N. B. Bell, and A. M. Abdullah, "Impact of ambient air pollution on locally grown rice cultivars (Oryza sativa L.) in Malaysia," *Water. Air. Soil Pollut.*, vol. 154, no. 1–4, pp. 187–201, 2004.

[14] R. Afroz, M. N. Hassan, and N. A. Ibrahim, "Review of air pollution and health impacts in Malaysia," *Environ. Res.*, vol. 92, no. 2, pp. 71–77, 2003.

[15] J. W. Koo, S. W. Wong, G. Selvachandran, H. V. Long, and others, "Prediction of Air Pollution Index in Kuala Lumpur using fuzzy time series and statistical models," *Air Qual. Atmos. Heal.*, vol. 13, pp. 77–88, 2020.

[16] F. Ahamad, M. T. Latif, M. F. Yusoff, M. F. Khan, and L. Juneng, "So near yet so different: Surface ozone at three sites in Malaysia," in *IOP Conference Series: Earth and Environmental Science*, vol. 228, pp. 1-5, 2019.

[17] N. D. A. Halim, *et al.*, "The long-term assessment of air quality on an island in Malaysia," *Heliyon*, vol. 4, no. 12, 2018.

[18] J. C. H. Wui, C. F. Pien, S. K. S. Kai, and J. SENTIAN, "Variability of the PM 10 Concentration in the Urban Atmosphere of Sabah and Its Responses to Diurnal and Weekly Changes of CO, NO_2, SO_2 and Ozone," *Asian J. Atmos. Environ.*, vol. 12, no. 2, pp. 109–126, 2018.

[19] H. I. Ahmad, A. Mustapha, H. Juahir, M. Alhaji, and H. Adamu, "Analysis of seasonal levels of atmospheric pollution in Terengganu, Malaysia," *FUDMA J. Sci. 2616-1370*, vol. 2, no. 2, pp. 88–100, 2018.

[20] A. Afzali, M. Rashid, M. Afzali, and V. Younesi, "Prediction of air pollutants concentrations from multiple sources using AERMOD coupled with WRF prognostic model," *J. Clean. Prod.*, vol. 166, pp. 1216–1225, 2017.

[21] Y. Alyousifi, N. Masseran, and K. Ibrahim, "Modeling the stochastic dependence of air pollution index data," *Stoch. Environ. Res. risk Assess.*, vol. 32, no. 6, pp. 1603-1611, 2017.

[22] A. Sulaiman, *et al.*, "Distribution ozone concentration in Klang Valley using GIS approaches," in *Journal of Physics: Conference Series*, vol. 852, 2017.

[23] N. Amil, M. T. Latif, M. F. Khan, and M. Mohamad, "Seasonal variability of PM 2.5 composition and sources in the Klang Valley urban-industrial environment," *Atmos. Chem. Phys.*, vol. 16, no. 8, pp. 5357–5381, 2016.

[24] H. A. Isiyaka and A. Azid, "Air quality pattern assessment in Malaysia using multivariate techniques," *Malaysian J. Anal. Sci.*, vol. 19, no. 5, pp. 966–978, 2015.

[25] S. R. A. Rahman, S. N. S. Ismail, M. F. Raml, M. T. Latif, E. Z. Abidin, and S. M. Praveena, "The assessment of ambient air pollution trend in Klang Valley, Malaysia," *World Environ.*, vol. 5, no. 1, pp. 1–11, 2015.

[26] N. A. Mabahwi, O. L. H. Leh, and D. Omar, "Urban air quality and human health effects in Selangor, Malaysia," *Procedia-Social Behav. Sci.*, vol. 170, pp. 282–291, 2015.

[27] M. Sahani, *et al.*, "A case-crossover analysis of forest fire haze events and mortality in Malaysia," *Atmos. Environ.*, vol. 96, pp. 257–265, 2014.

[28] A. Azid, *et al.*, "Prediction of the level of air pollution using principal component analysis and artificial neural network techniques: a case study in Malaysia," *Water, Air, Soil Pollut.*, vol. 225, no. 8, pp. 1-14, 2014.

[29] A. Azid, *et al.*, "Spatial analysis of the air pollutant index in the Southern Region of Peninsular Malaysia using Environmetric Techniques," in *From Sources to Solution*, Springer, pp. 307–312, 2014.

[30] J. Othman, M. Sahani, M. Mahmud, and M. K. S. Ahmad, "Transboundary smoke haze pollution in Malaysia: Inpatient health impacts and economic valuation," *Environ. Pollut.*, vol. 189, pp. 194–201, 2014.

[31] S. N. S. A. Mutalib, *et al.*, "Spatial and temporal air quality pattern recognition using environmetric techniques: a case study in Malaysia," *Environ. Sci. Process. Impacts*, vol. 15, no. 9, pp. 1717–1728, 2013.

[32] N. R. Awang, N. A. Ramli, N. I. Mohammed, and A. S. Yahaya, "Time series evaluation of ozone concentrations in Malaysia based on location of monitoring stations," *Int. J. Eng. Technol.*, vol. 3, no. 3, pp. 390–394, 2013.

[33] D. Dominick, M. T. Latif, H. Juahir, A. Z. Aris, S. M. Zain, "An assessment of influence of meteorological factors on PM10 and NO2 at selected stations in Malaysia," *Sustain. Environ. Res.*, vol. 22, no. 5, pp. 305–315, 2012.

[34] A. Makmom Abdullah, M. Armi Abu Samah, and T. Yee Jun, "An overview of the air pollution trend in Klang Valley, Malaysia," *Open Environ. Sci.*, vol. 6, no. 1, pp. 13-19, 2012.
[35] M. Bin Hitam and H. B. Borhan, "FDI, growth and the environment: Impact on quality of life in Malaysia," *Procedia-Social Behav. Sci.*, vol. 50, pp. 333–342, 2012.
[36] M. T. Latif, L. S. Huey, and L. Juneng, "Variations of surface ozone concentration across the Klang Valley, Malaysia," *Atmos. Environ.*, vol. 61, pp. 434–445, 2012.
[37] D. Dominick, H. Juahir, M. T. Latif, S. M. Zain, and A. Z. Aris, "Spatial assessment of air quality patterns in Malaysia using multivariate analysis," *Atmos. Environ.*, vol. 60, pp. 172–181, 2012.
[38] L. Juneng, M. T. Latif, and F. Tangang, "Factors influencing the variations of PM10 aerosol dust in Klang Valley, Malaysia during the summer," *Atmos. Environ.*, vol. 45, no. 26, pp. 4370–4378, 2011.
[39] M. T. Latif, *et al.*, "The impact of urban growth on regional air quality surrounding the Langat River Basin, Malaysia," *Environmentalist*, vol. 31, no. 3, pp. 315-324, 2011.
[40] S. Z. Azmi, M. T. Latif, A. S. Ismail, L. Juneng, and A. A. Jemain, "Trend and status of air quality at three different monitoring stations in the Klang Valley, Malaysia," *Air Qual. Atmos. Heal.*, vol. 3, no. 1, pp. 53–64, 2010.
[41] "Met One Instrument Distributors," Met One Instrument Distributors. [Online]. Available: https://metone.com/distributors/. [Accessed on Mar 26, 2020].

BIOGRAPHIES OF AUTHORS

Raja Sher Afgun Usmani received the B.S. degree in computer science from International Islamic University, Islamabad, Pakistan, in 2011 and the M.S. degree in computer science from International Islamic University, Islamabad, Pakistan, in 2017. He is currently pursuing the Ph.D. degree in computer science at Taylor's University, Malaysia. From 2010 to 2015, he worked with various software development companies as a software developer in Islamabad, Pakistan. From 2015 to 2018, he was a Senior Lab Engineer with the International Islamic University, Islamabad, Pakistan. His research interest includes the Geographical Information Systems, Spatial Data, Big Data, Data Mining and Data Science.

Ms. Wan Nurul Farah Wan Azmi is a Researcher in the Environmental Health Research Centre, Institute for Medical Research (IMR), Malaysia. She has joined IMR since 2011. She obtained her Bachelor of Science with Honours in Environmental Science (Environmental Chemistry) and Master of Science (Environmental Assessment and Monitoring) from National University of Malaysia (Universiti Kebangsaan Malaysia). She is currently pursuing a PhD in Science at Taylor's University, Malaysia. Her research mainly focused on air and water contamination in the environment and its risk towards human health.

Akibu Mahmoud Abdullahi received the B.A. degree in Arabic Language from Bayero University Kano, Nigeria, in 2011. B.S. degree in Information Technology (It) from Almadinah International University, Selangor, Malaysia, in 2016 and the M.S. degree in Instructional Multimedia from University Sains Malaysia (USM), Penang, Malaysia, in 2017. He is currently pursuing the Ph.D. degree in computer science at Taylor's University, Malaysia. From 2016 to 2018, he was as IT Help Desk Technician with Labtech International Limited, Malaysia. His research interest includes the Learning Analytics, Big Data, Educational Data Mining and Data Science.

Ibrahim Abaker Targio Hashem has received his Master degree in Computer Science from the University of Wales, Newport and Doctor of Philosophy (Ph.D.) degree in Computer Science from University of Malaya. Dr. Hashem obtained professional certificates from CISCO (CCNP, CCNA, and CCNA Security) and APMG Group (PRINCE2 Foundation, ITIL v3 Foundation, and OBASHI Foundation). He is presently working as a lecturer at the Department of Computing and IT, Taylor's University, Selangor, Malaysia. He has published a number of research articles in refereed international journals and magazines. His numerous research articles are very famous and among the most downloaded in top journals. His area of interest includes Big Data, Cloud Computing, Distributed Computing, and Machine Learning. He is an active member of Mobile Cloud Computing center, Malaysia.

Thulasyammal Ramiah Pillai is a committed lecturer with over 25 years of experience at higher academic institutions and have taught students from various social and cultural backgrounds. Her research interests include Time series analysis, Probability and Statistics, Statistical Modelling and Signal Processing. Her current research project is "Modelling and visualization of air pollution and its impact on health".