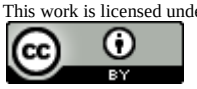
Available online at: https://ijact.in

Date of Submission	13/08/2020
Date of Acceptance	17/09/2020
Date of Publication	30/09/2020
Page numbers	3831-3840 (10 Pages)



ISSN:2320-0790

An International Journal of Advanced Computer Technology

AIR QUALITY INDEX USING MACHINE LEARNING – A JORDAN CASE STUDY

Khalid M.O.Nahar[1], Mohammad Ashraf Ottom[2], Fayha Alshibli[3] and Mohammed M. Abu Shquier[4]

[1]Department of Computer Sciences, Yarmouk University, Irbid, Jordankhalids@yu.edu.jo
[2]Department of Information Systems, Yarmouk University, Irbid, Jordanottom.ma@yu.edu.jo
[3]Department of Land-Water and Environment, Jordan University, f.shibli@ju.edu.jo
[4]Department of Computer Sciences, Jerash University, shquier@jpu.edu.jo

ABSTRACT: Predicting changes in air pollutant concentrations due to human and nature drivers are critical and challenging, particularly in areas with scant data inputs and high variability of parameters. This paper builds an Air Quality Index (AQI) model using Machine Learning algorithms and techniques. The paper employs Machine Learning Algorithms such as Decision Tree (DT), Support Vector Machine (SVM), k-Nearest Neighbor (k-NN), Random Forest (RF) and Logistic Regression. The model can predict the most pollutant factors from real readings published daily by the Jordan Ministry of Environment (MoEnv) for the period from January 2017 to April 2019. Jordan has prioritized air quality problems by establishing detection and monitoring stations in 12 positions across the country to measure Air Quality (AQ). Pollutant concentrations recorded by MoEnv use fully forewarn official organizations and individuals of daily air quality in the atmosphere over time and beneficially used by health and climate studies organizations. The study has detected the most contaminated sites and determined the pollutant concentrations. These estimates will indicate the most influenced pollutants and their behavior in the pollution process for further recommendations and actions to effects cardiopulmonary patients, environmental and climate researches, as well as to vulnerable ecosystems.

Keywords*:* Machine Learning, Air Pollution, Air Quality.

I. INTRODUCTION

Immoderate amounts of gases, particles, or molecules that brought into Earth's atmosphere certainly cause Air Pollution (AP). These amounts are sourced from: anthropogenic and technological actions such as combustion and industrial processes, or natural actions like woodfires and volcanic activity. The consequences of high pollutants' concentrations have aroused interest in eliminating its sources and mitigating its effects. Human health, climate, and ecosystems are the most likely liable to the effects of air pollution. Many studies have long prompted to health implications for air low quality leading to cardiopulmonary health issues, genetic childhood asthma [1] and more recently to substantially effects on neurological problems [2]. Because of its large-scale effects on ecosystems and its cross-boundary nature, environmental scientists are increasingly required to meet situations calling for scientific data to resolve environmental problems. Starting from the contaminants release mechanisms, characterizing the risks of carbon dioxide and greenhouse emissions on vegetation, animals, atmosphere, water bodies, soils, and long-term effects on warming the globe urge scientists to speculate the concentrations of air pollutants by the time [3]. The increasing greenhouse gas concentrations will lead to higher emission levels, and consequently, higher trapping of radiations in the atmosphere which causes warming up the earth's temperature. Climate models have projected the global mean annual greenhouse gas emissions which are; CO2, CH4, and NO_2 as 538, 1580, and 372 Parts Per Million (ppm) respectively will increase radiations to 4.5 Watt per Square Meter (w/m^2) by 2100, in contrast, if the concentrations reach 670, 1650, and 406 ppm by 2100, it will lead to 6 W/m^2 of radiative forcing [4]. Regional climate models have projected

the total monthly Sulfur Oxide (SO_2) emissions at ground levels of Amman city which leads to increase the temperature by one-degree Celsius and the radiations by 3.0 W/m^2. The highest monthly emission of SO_2 was 88.9 ng/m^2/s during year 2000. It gives a highly-confidence scientific proof of air pollutant's impacts on increasing the average temperature of the earth. This study focused on diffuse outdoor pollutants in Jordan which mainly are: Particulate Matter 10µm (PM_{10}), Nitrogen Peroxides (NO_2), Sulfur Dioxide (SO_2), Carbon Monoxide (CO), Ozone (O_3), and Hydrogen Sulfide (H_2S). Datasets are provided by the Jordanian Ministry of Environment [5].

Potential air pollution impacts are estimated using air quality computational simulation models. A model requires a sufficient amount of data at specific receptor locations and times. Machine Learning is one of the successful sciences that have been deployed recently in many applications due to recent advancements computing technologies and the availability of data, which added many benefits in various fields, including healthcare, finance, retails, and environment. Accurate predictions and classifications in Machine learning projects depend on several measures such as data quality, therefore, biased, low quality, or insufficient dataset can cause low and unjustified accuracy. Many organizations and websites provide publicly online datasets for research purposes such as governmental organizations and websites (Kaggle and UCI Machine Learning Repository), however, these sites are inappropriate for all projects like AQI due to the variability of meteorological data and pollution sources in the area that collected in various spatial and temporary locations [6], [7].

This work employs variant machine learning algorithms rather than reliance on one algorithm to achieve a better computer model predicting the most current actual pollutant factors affect the climate, with sufficient and useful performance and accuracy. Model inputs were obtained from MoEnv observations across 12 sites, processed and filtered for higher model performance and optimization. Different algorithms were run to predict the pollutant concentrations and then each algorithm was assessed for higher performance. The study has detected the most contaminated sites over the research period and determined the concentration of each pollutant. In turn, this will help to project the pollution source, eliminate the concentration, and mitigate the effects on different aspects.

II. LITERATURE REVIEW

The total of Carbon Monoxide, Suspended Particulate Matter and Sulphur Dioxide levels in the suburban and city of Amman are higher than international standards, according to a study [8] based on data given by MoEnv. Another research [6] investigated machine learning algorithms predicting air quality index in Czech, it aimed to design computer models for predicting AQIs which allow modeling complex and non-linear processes within the formation frame. Researchers showed that the models are capable of: (1) learning relations between pollutants; (2) applying them later on real data; (3) and finally processing uncertainty related to measuring both APs and meteorological variables. A study using machine learning to classify the Air Quality Level in Beijing city was conducted by [9], the author applied Random Forest, Support Vector Machine, and ranking methods to determine the top pollutants such as CO, $PM_{2.5}$ (fine inhalable particles, with diameters that are generally 2.5 micrometers and smaller), and PM_{10} [inhalable particles, with diameters that are generally 10 micrometers (µm) and smaller]. The experiment produced about 95% accuracy when using the SVM algorithm. A recent study[10] used several machine learning algorithms to prove the ability of computing methods determining air pollutants index, the authors used four different algorithms: Neural Network (NN), k-Nearest Neighbors algorithm (kNN), Decision Tree (DT) and SVM. The conducted experiment was performed on a dataset collected from the state of Macedonia in 2017; they found that NN produced better accuracy of 0.92, KNN and SVM approximately 0.8, and DT 0.78. In another research [10], Machine learning algorithms were used to predict the hourly spot concentration pollutants indexes such as O_3, PM2.5, and nitrogen dioxide (NO_2) for the coming 48 hours for several locations in Canada. Data were obtained from an air quality model (GEM-MACH15). Different machine learning algorithms were used such as multiple linear regression (MLR), multilayer perceptron neural networks (MLP NN), extreme learning machine (ELM). The study has shown the potential of using machine learning methods to improve air quality forecasts.

Table 1: Selected Previous Studies on Air Pollutants Forecast

Authors / Year	Goal	Algorithms used	Dataset used	Results / Accuracy
[11] (Peng, 2015)	Building forecast models for pollutant concentrations	multiple linear regression (MLR) multilayer perceptron neural networks (MLP NN).	Canada dataset	For each factor (O_3, $PM_{2.5}$, NO_2), the algorithms used outputted different accuracy percentage Accuracy: 75%
[6](Hajek &Olej, 2015)	Design a model for predicting AQIs	TSFIS, RBF and MLP neural networks and support	Data measured in Dukla, Rosice, and Brnenska in	Designing two models for prediction of AQI^{t+1} using computational methods.

		vector regression (SVR)	the Czech Republic	Dividing their AQI to 6 categories Accuracy: N.A
[9](Yang & de Loera, 2018)	Applying Random Forest (RF) to regress the Air Quality Index and Support Vector Machine to classify the Air Quality Level	Random Forest (RF) Support Vector Machine	Data is collected from China National Urban Air Quality Real-time Publishing Platform	The RF method gives a 99 % prediction accuracy For the 3 kernels (radial, polynomial and linear) the classification accuracy on the test data are: (0.9377155, 0.6090913, 0.9343915)
[10] (Veljanovskal&Dimoski, 2018)	Comparing between 4 different algorithms to check the best performance and accuracy for predicting the AQI	k-nearest neighbor (k-NN) Support Vector Machines (SVM) Decision Tree (DT) Neural Network (NN)	Macedonia Dataset	• NN accuracy: 92.3% • KNN: 80.0% • DT: 78.0% • SVM: 80.0%
[12] (Zhu et al., 2018)	Using machine learning approaches to predict the hourly concentration of Air Pollutants	multi-task learning (MTL) ASSG and SSG methods	Chicago area.	Developing efficient machine learning methods for Air Pollutant prediction Accuracy: N.A
[13](Soh et al., 2018)	Forecasting air quality for up to 48 h using a combination of multiple neural networks, to extract spatial-temporal relations.	artificial neural network a convolutional neural network, and a long-short-term memory	Taiwan and Beijing data sets	Proposed an Air Quality forecasting system using data-driven models, ST-DNN to predict PM2.5 over 48 hours. Accuracy: N.A

A recent study [13] has focused on a small diameter of fine particulate matter (PM2.5) as one of the popular pollutants indexes that cause several diseases such as cardiovascular disease. The study intended to predict air quality index for every 48 hours in Taiwan and Beijing. The study used neural networks, such as traditional artificial neural networks and deep learning approach (convolutional neural network) to extract knowledge and interrelation between spatial-temporal data. The model used historical meteorology data hourly-basis across different stations and information related to the elevation space to examine the effects of topography on air quality. Experiments have shown the model ability to utilize weather data and machine learning to predict the behavior of the coming AQI Research has shown the use of metrology data for indoor air quality index, which brought the attention for the use of pollutants index inside the home to bring comfy, clean-air and healthy environment indoors and homes[14]. The research used some major pollutants such as PM_{10}, O_3, $PM_{2.5}$, CO_2, CO, HCHO, TVOC, SO_2, NO_2, fungi, and bacteria in Taiwan. The study conducted different algorithms that might enhance pollutant's index prediction accuracy for indoors and homes.

III. AIR QUALITY IN JORDAN

Jordan population is approximately ten million according to the Jordanian Department of Statistics, covering an area of Approximately 89,000 km^2, which has undergone a rapid development of different industries resulting in the contaminated air quality. The prospective air pollutants in the city (Amman) for instance, are collected from different sources and sinks that mainly are: the high sulfur content of heavy fuel oils (typically 3 percent Sulfur), the open-air incineration of domestic wastes (approx. 600 tones/day), the existence of sand and lime quarries in and around the city and the burning of used lubricating oils by bakeries, smelters, pottery factories and other small-scale industries located in residential areas and cities[8].

The Ministry of Environment is monitoring the ambient air quality of 12 areas around Jordan and keeps the records in a dataset. The 12 areas are distributed in Amman, Al-Zarqa city, and Irbid as 7, 3, and 2 sites respectively [5]. The reporting of AQ is functioning by the Ministry of Environment per Article (4) of the Environmental Protection Act No. 52 of 2006.

Ambient air quality limits in Jordan are recommended by the Ambient Air Quality Standards (No. 1140/2206). A summary of these limits is presented in Table 2[5].

Table 2: Jordanian Ambient AQ Standards[5]

Air Pollutant	Average Time	Maximum Allowable Concentration in the Ambient Air	Number of Allowed Exceedances
Sulfur Dioxide	1 Hour	0.30 mg/kg	3 times within a given month in one

(SO2)			year
	24 Hour	0.14 mg/kg	Once a year
	1 Year	0.04 mg/kg	--
Carbon Monoxide (CO)	1 Hour	26 mg/kg	3 times within a given month in one year
	8 Hour	9 mg/kg	3 times within a given month in one year
Nitrogen Dioxide (NO_2)	1 Hour	0.21 mg/kg	3 times within a given month in one year
	24 Hour	0.08 mg/kg	3 times within a given month in one year
	1 Year	0.05 mg/kg	--
Hydrogen Sulfide (H_2S)	1 Hour	0.03 mg/kg	3 times within a given month in one year
	24 Hour	0.01 mg/kg	3 times within a given month in one year
Ozone (O3)	1 Hour	0.08 mg/kg	--
	8 Hour	0.12 mg/kg	--
PM_{10}	24 Hour	120 µg/m3	3 times within a given month in one year
	1 Year	70 µg/m3	--

IV. DATASET ACQUISITION

Ministry of Environment publishes a daily report of the six pollutant readings via the Ministry's official website on a daily PDF format file. Each report has a list of all monitoring stations as shown in Figure 1.

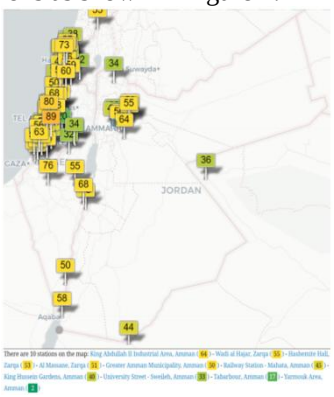


Figure 1: List of Air detection stations distributed in Jordan[1]

Furthermore, charts for daily pollutants averages are introduced specifying the limits according to the Jordanian Ambient Air Quality Standards (1140/2006) as shown in Figure 2 and Figure 3, detailing the measurements of the PM10 µm and NO_2 in all stations.

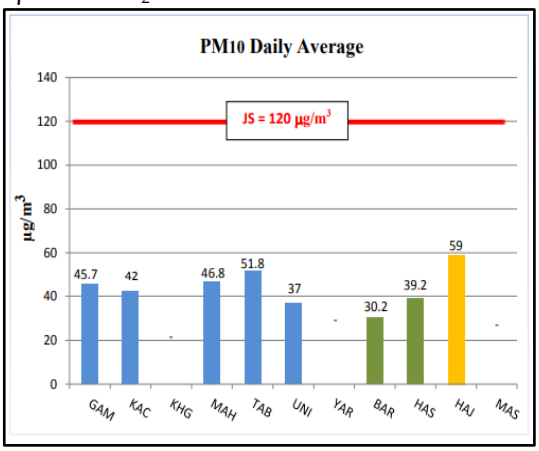


Figure 2: Measurements of PM_{10} for the 13 stations [19]

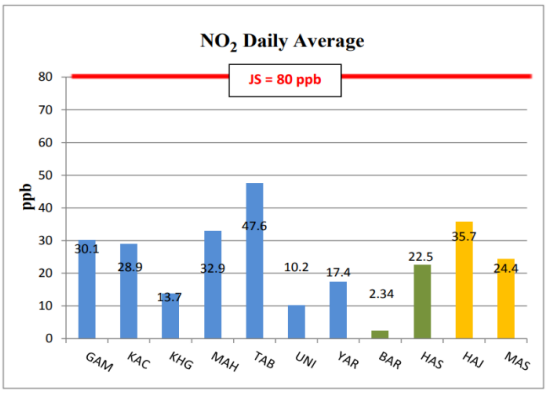


Figure 3: Measurements of recorded NO_2in the 13 stations [19]

Meteorological observations are also recorded, including Ambient Relative Humidity, Ambient Temperature, Average Wind Direction, and Average Wind Speed (Figure 4).

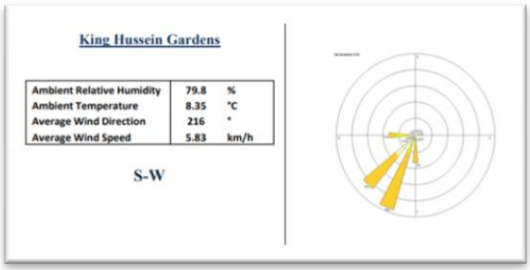


Figure 4: Climate dataset for Gardens area in Amman [19]

A sample reading for the period of 24 hours for all areas (stations) are shown Table 3.

[1] Taken from https://aqicn.org/map/amman/, on 21_9_2020 at time 12:55 AM.

Table 3: Sample data daily reading in ppb (parts per billion) [19]

	pollutants		PM10	NO_2	SO_2	CO	O3	H_2S
Jordanian standard (JS) 2006/1140			**120 µg/m3**	**80 ppb**	**140 ppb**	**9000 ppb**	**80 ppb**	**10 ppb**
	Station	**Abr**	24 HR AVG	24 HR AVG	24 HR AVG	8 HR AVG MAX/24 HRS	8 HR AVG MAX/24 HRS	24 HR AVG
1	Greater Amman Municipality	GAM	**45.7**	**23.8**	**5.51**	**2920**	**None**	**None**
2	King Abdullah II/Industrial City / Sahab	KAC	**42**	**19**	**8.1**	**None**	**None**	**None**
3	King Hussein Gardens	KHG	**-**	**11.4**	**3.55**	**None**	**41.6**	**None**
4	Marka – Mahata	MAH	**46.8**	**30.9**	**10**	**None**	**None**	**None**
5	Northern Bus Station Tabarbour	TAB	**51.8**	**45.7**	**None**	**2057**	**None**	**None**
6	University street Sweileh	UNI	**37**	**9.17**	**None**	**None**	**None**	**None**
7	Wadi Rimam Yarmulke Garden	YAR	**-**	**3.04**	**2.66**	**None**	**None**	**None**
8	Al Barha street	BAR	**30.2**	**2.77**	**5.34**	**None**	**52.7**	**None**
9	AL Hassan Sport City	HAS	**39.2**	**21.8**	**None**	**1826**	**None**	**None**
10	Health Center Wadi Hajjar	HAJ	**59**	**45**	**25.6**	**1442**	**None**	**None**
11	Main slaughter house/Masane' Zone	MAS	**-**	**23.3**	**9.83**	**None**	**None**	**None**
12	Arab Bank Garden	ANB	**None**	**None**	**-**	**None**	**None**	**-**
13	Um sharbak	UMS	**None**	**None**	**-**	**None**	**None**	**-**

Jordanian Ministry of Environment (MoEnv) data requires preprocessing and to be saved into spreadsheet document (Excel) as shown in figure 5.

	A	B	C	D	E	F	G	H	I	J	K
1	year	No	Station	PM10	NO2	SO2	CO	O3	H2S	Pollution	Pollution1
2	2017	1	GAM	101	34.3	15.5	2298	0	0	0	N.P
3	2017	2	KAC	79.8	13.4	5.89	0	10.1	0	0	N.P
4	2017	3	KHG	72.2	20.7	13.1	0	20	0	0	N.P
5	2017	4	MAH	118	23.2	9.18	0	0	0	0	N.P
6	2017	5	TAB	89	30.6	0	1522	0	0	0	N.P
7	2017	6	UNI	85.4	9.65	0	0	0	0	0	N.P
8	2017	7	YAR	95	18.6	21.9	0	0	0	0	N.P
9	2017	8	BAR	66.2	28.9	32.6	0	17.1	0	0	N.P
10	2017	9	HAS	90.3	20.2	0	5428	0	0	0	N.P
11	2017	10	HAJ	98.7	15.7	7.9	2533	0	0	0	N.P
12	2017	11	MAS	126	12.2	2.77	0	0	0	1	PM10
13	2017	12	ANB	0	0	8	0	0	2.68	0	N.P
14	2017	13	UMS	0	0	5.38	0	0	1.32	0	N.P

All data | Deploy data | Train and Validation data

Figure 5: Sample of the dataset Excel file

A. *Data Cleaning:* the dataset required preparation and preprocessing before the actual modeling process, consequently. Preparation of data includes filling the missing day/station readings with the average value between the previous and next day readings. The missing readings which were reported as 'None' have been set to zero values.

B. *Features Selection Process:* MoEnv records the daily pollutant factors conformed to the Jordanian Standard (1140/2006). One extra attribute column has been added to the Excel dataset that includes a binary (0, 1) classification (The Area is Polluted or Not Polluted). Another column was added to represent the name of the most effective pollutant, and so the classification values. In addition to, labeling one of the six pollutants or No pollution. The features are independent variables, and all affect the dependent variables (class label) which are (pollution and pollution1 columns). The first three columns were not used in the machine learning process due to their ineffectiveness.

V. METHODOLOGY

The available dataset was chosen for the period between 1/1/2017 and 30/4/2019 across 13 different sites in Jordan, which consists of 9777 records and divided into 9330 records for Train and Validation while the remaining 560 records were used for outlet testing. Different factors and compounds were observed and considered as pollutants. Pollution statistics in the Train and Verification portion of the dataset indicated in the table below and the remaining days were reported clean (No pollution), see Figure 6.

Ser.	Factor	Occurrence
1	PM10	757
2	NO2	4
3	SO2	2
4	CO	10
5	O3	1
6	H2S	8

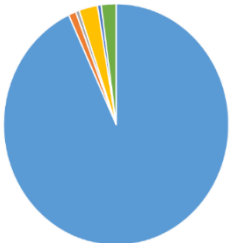


Figure 6: Pollution statistics

The purpose of this study is to build several classification models for predicting the Air Pollution Factors for a given day based on the baseline historical climate dataset. After data preparation, different Machine Learning algorithms were used: Decision Tree (DT), Support Vector Machine (SVM), k-nearest neighbor (k-NN), Random Forest (RF), and Logistic regression algorithms. Data Modeling software and packages, such as Knime and Orange, were used to build classification models. The Knime and Orange were elaborated to apply the following tasks (as shown in Figure 7): (1) Loading dataset, (2) Partitioning dataset (70% training and 30% Verifying), (3) model building process, (4) Writing the model to HD storage, (5) Verifying/testing (predictor), (6) Showing the score of results.

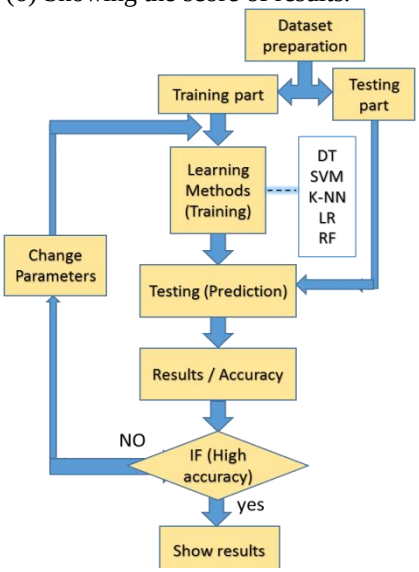


Figure 7: Work flowchart

VI. ALGORITHM DEFINITION AND PERFORMANCE

Decision tree (DT): The decision tree is considered among the most effective and common algorithms for classification and future prediction. DT is a conceptual tree alike model, where each internal node represents a feature that best split the data into subsets using statistical measures such as information gain and gain ration, the process of splitting data is a recursive process until reaching the leaf (normally one of class labels). In machine learning, DT is one of the supervised learning algorithms. They feature variable contains 7 choices (6 pollutant factors and 7th is labeled "No pollution" which is compatible with the study goal to estimate the most influence pollution factor (one of the six) or (No Pollution). When the Decision tree was applied, the accuracy reflected 99.928 % for the medium tree as shown in Figures 9 below.

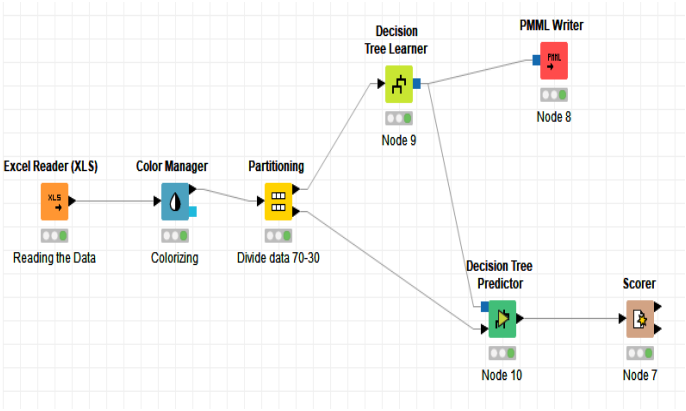


Figure 8: Decision Tree Learner workflow

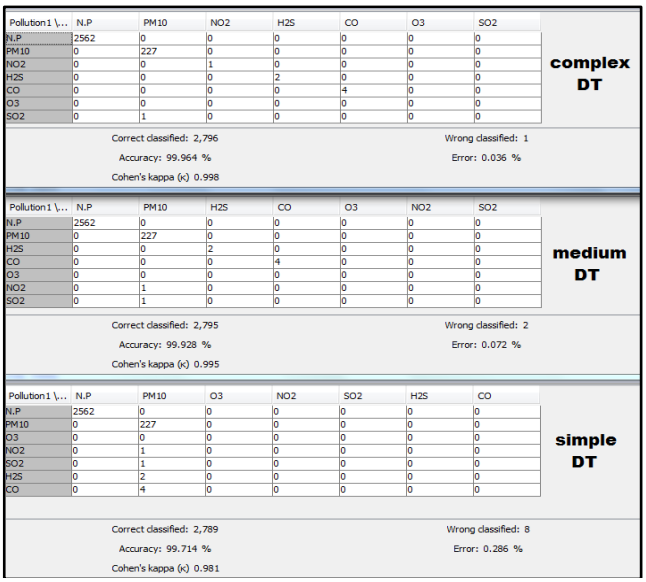


Figure 9: Decision Tree Learner Confusion Matrix Scorer

- **Support Vector Machine (SVM)**is a well-known machine learning algorithmfor classification and prediction purposes. SVM label each instanceto a certain and given target class, by making it a non-probabilistic binary linear classifier. The model will focuson instances at the edge of each clusters and use the middle point between clusters as threshold, thenallocate each new instancein accordance to its distance to threshold, the distance between the edge of clustered instances and threshold is called margin[13-18]. The used dataset was typical for using SVM. The accuracy here was 99.837% using orange software.

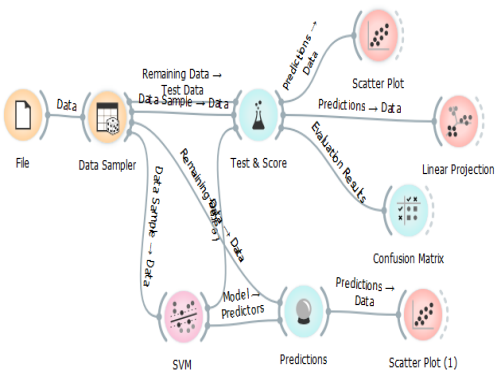


Figure 10: SVM Learner workflow

A tabular in Figure 11 represents the confusion matrix of the SVM model using orange software; 10 values out of 2765 (testing values) were considered as error.

Predicted

Actual	CO	H2S	N.P	NO2	PM10	SO2	Σ
CO	0	0	0	0	6	0	6
H2S	0	3	0	0	0	0	3
N.P	0	0	2531	0	0	0	2531
NO2	0	0	0	0	2	0	2
PM10	0	0	0	0	221	0	221
SO2	0	2	0	0	0	0	2
Σ	0	5	2531	0	229	0	2765

Figure 11: SVM Learner Confusion Matrix Scorer

- **k-nearest neighbor (k-NN)** is an example of lazy learning algorithm used for classification and regression. k-NN classify instances based on the distance between a given instance and other instances in the training dataset, distance measures used are Euclidean and Manhattan[14]. Figure 12 is the pictorial view of KNIME / k-NN work flow model.

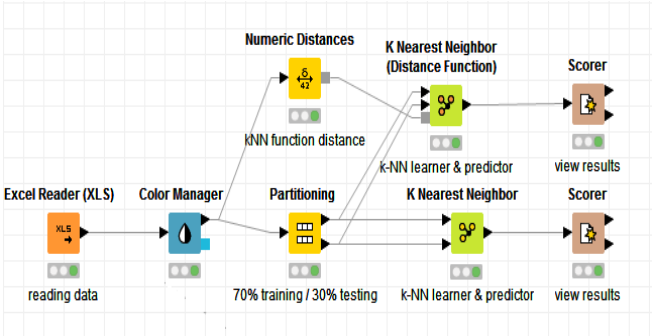


Figure 12: k-NN Learner workflow

When using the k-NN with k=6, the best accuracy was 93.886 %, while using k-NN with the numerical distance function, the accuracy was approaching up to 99.714 % as shown in a sample Figure 13.

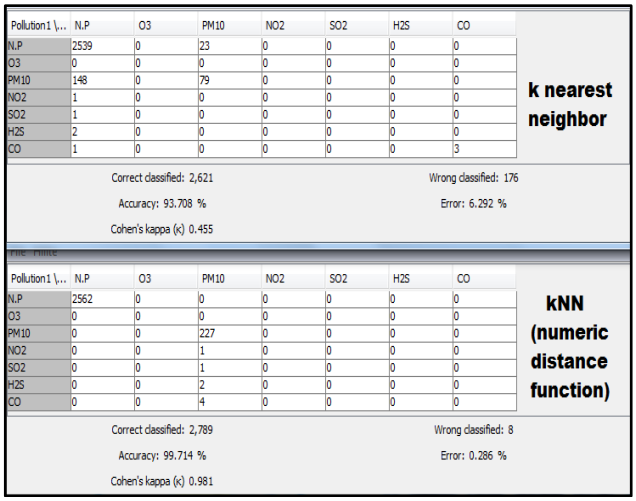


Figure 13: k-NN Learner Confusion Matrix Scorer

- **Random Forest (RF)** is one of the most popular machine learning algorithms for regression and classification tasks. RF creates a number of decision trees called forest trees to enhance the prediction process and produce higher accuracy. Building RF tree is similar to decision tree (DT) using information gain or other measures. Since RF is a set of DTs; each tree obtains a certain output and RF will choose the majority output produced by DTs or the mean in case of regression problem [15] RF is used over DT because of its ability to handle missing and solve the overheating problem. Accuracy using this type was 99.714 %

- **Logistic Regression:** is a popular statistical machine learning algorithm for classification problems, the prediction of output is performed using nonlinear function such as sigmoid and logit functions[16]. Accuracy resulted from this type was 91.598 %

VII. MODELS COMPARISON AND DISCUSSION

Table (4) shows the comparison of the final results between all previous Machine Learning algorithms used in this study: DT, SVM, k-NN, RF, and Logistic Regression. The study was conducted using a dataset from Jordan across 13 different stations in the country. All algorithms resulted in good accuracy while DT had the highest performance.

Table 4: Final results and accuracy

Algorithms	Accuracy	
Decision Tree (DT)	Complex	99.964 %
	Medium	99.928 %
	Simple	99.714 %
Support Vector Machine (SVM)	99.837 %	
k-nearest neighbor (k-NN)	k-NN (k=6)	93.708 %
	Numeric distance k-NN	99.714 %

Random Forest (RF).	99.714 %
Logistic regression.	91.598 %

When running DT and SVM, the performance was the highest scoring 99.71 to 99.96%, which is the same as using the k-NN method substituting k=6. Whereas, the accuracy resulted from running Logistic Regression and Numeric Distance k-NN were 91.6 and 93.71% respectively. The chart in Figure 14 represents the accuracies of the used algorithms. the least 91.6% accuracy was recorded.

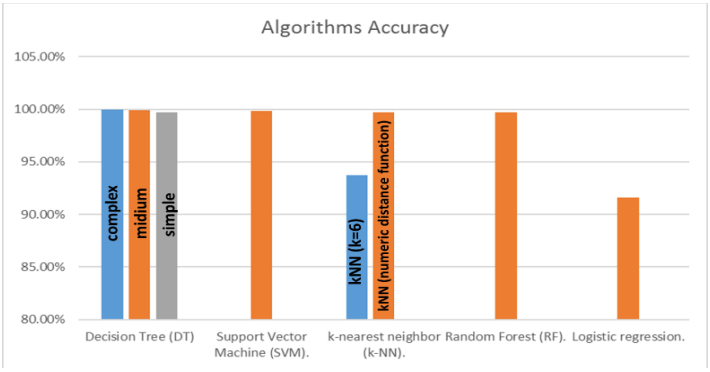


Figure 14: Algorithms Accuracy

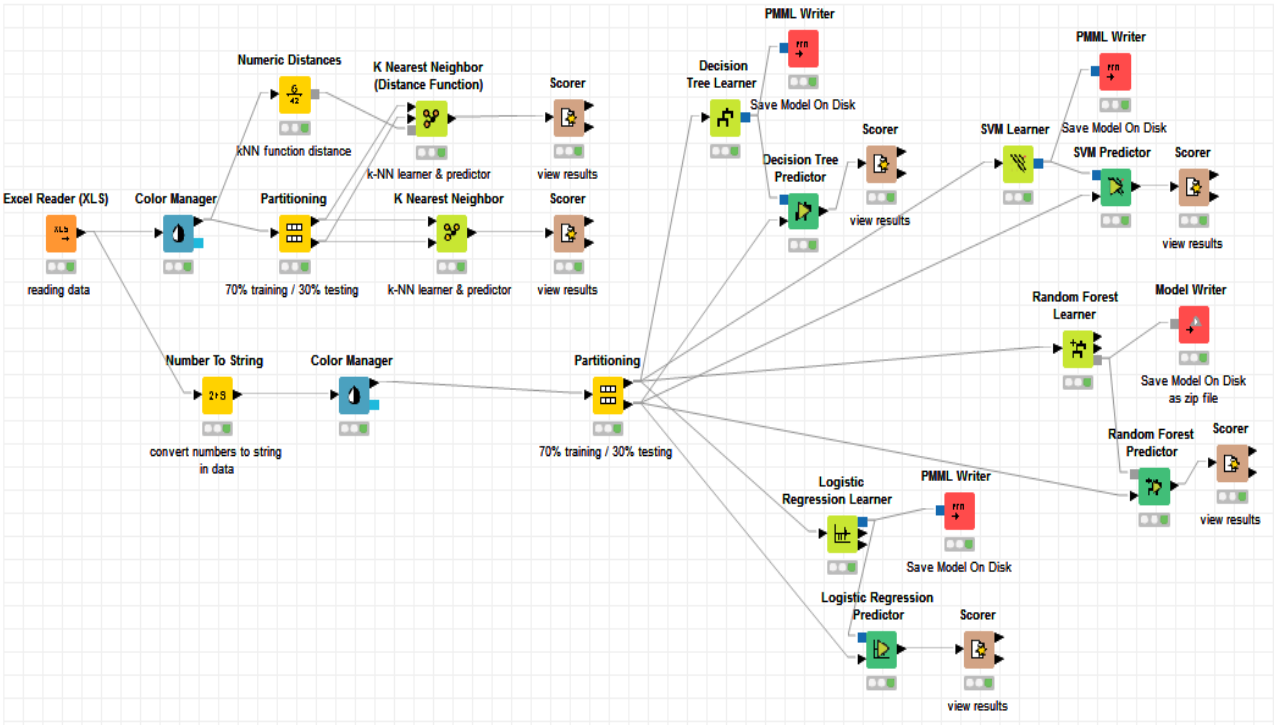


Figure 16: The Overall Schematic Methodology Flowchart (all models in one workflow); the left shows the non-parametric algorithm and the right half shows the parametric algorithms.

A. *Model Deployment and Outlet Testing:*

The models coming out of the study method were saved to the hard drive and then loaded in KNIME software again in the Deploy workflow illustrated in Figure 15. As mentioned earlier, 560 out of 9777 readings were used for outlet testing. The results showed 550 correct values and only 10 values were predicted pollution but using different factors. The overall algorithms employed in this study are briefed in one workflow in Figure 16.

B. *Pollution Index:*

The results showed the susceptible regions to highest PM_{10} pollution have witnessed during the study period; ranged from 120 to 548 ppb; MAS, KAC, MAH, HAJ, GAM, KHG, TAB, UNI, YAR, HAS & BAR stations. Whereas the most polluted records among stations due to NO_2 are MAH, producing more than 80.0 ppb. The main pollutant for BAR and YAR sites is

The amount SO_2 is (340-600 ppb). In contrast, H_2S was approaching higher than 55.0 ppb in UMS and less than 20.0 ppb in ANB stations. Carbon mono Oxide is not the only pollutant in the HAS site, reached more than 9000 ppb, but also NO_2 contaminated the air of the site with an amount of 4139 ppb. Eventually, the total contamination records were 874 marked as polluted among the 9776 observations. Predictions the number of pollutants in the air will help to mitigate the sources of contaminations and conserve the sinks which absorbed it.

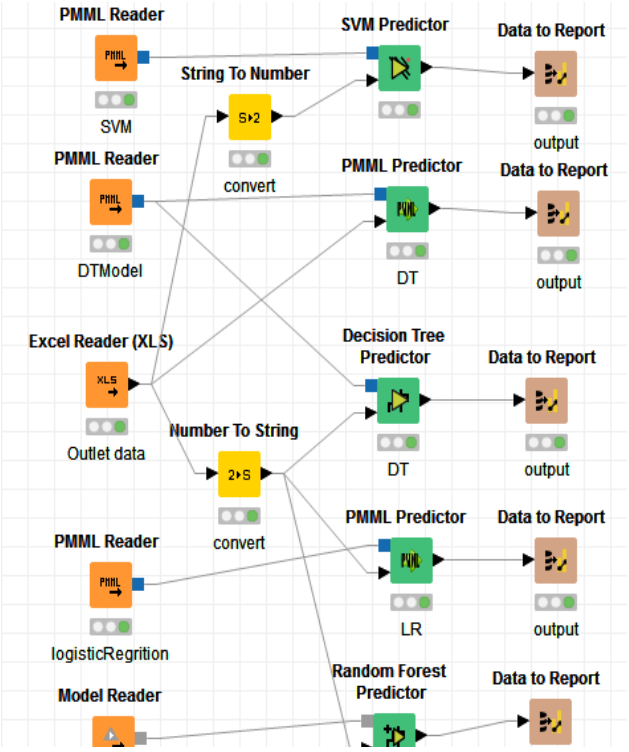


Figure 15: Deployment of models

VIII. Conclusion

Since the Air Quality and measuring the number of pollutants; and consequently, pollution depends on the identification of AQI in the source countries as each country has its management approach. Jordan has a method to calculate AQ which is not far from the other international methods. The AQ is classified into polluted area with specifying the most pollutant factor (PM_{10}, NO_2, SO_2, O_3, CO, H_2S) on a table. This study produced a model using the Machine Learning Algorithms: Data Tree (DT), Support Vector Machine (SVM), k-nearest neighbor (k-NN), Random Forest (RF), and Logistic Regression. The model is capable of predicting the most pollutant factor from daily observations published by the Ministry of Environment of Jordan scoring high accuracy not less than 92%.

For future work, the study recommends building a web-based interface using python or java that uses the PMML models produced from this study to predict AQI in Jordan. Data and Information available in this study would be beneficial for any study in the field of predicting AQI from time-series data as there were some days in the data published by MoENV with missing readings due to station error readings or instruments malfunctioning within the station.

IX. References

[1]. I. Ungváriet al., "Relationship between air pollution, NFE2L2 gene polymorphisms and childhood asthma in a Hungarian population," Journal of community genetics, vol. 3, no. 1, pp. 25–33, 2012.

[2]. J. Kotcher, E. Maibach, and W.T. Choi, "Fossil fuels are harming our brains: identifying key messages about the health effects of air pollution from fossil fuels," BMC public health, vol. 19, no. 1, p. 1079, 2019.

[3]. D. A. Vallero, Fundamentals of air pollution. Academic press, Massachusetts, 2014.

[4]. D. P. van Vuuren et al., "The representative concentration pathways: an overview," Climatic change, vol. 109, no. 1–2, p. 5, 2011.

[5]. MoEnv, "Annual Air Quality Report," Amman, 2017.

[6]. P. Hajek and V. Olej, "Predicting common air quality index--The case of Czech Microregions," Aerosol and Air Quality Research, vol. 15, no. 2, pp. 544–555, 2015.

[7]. P. Ghosh, "The Impact of Data Quality in the Machine Learning Era," DataVersity, 2018.

[8]. A. A. Al-Hasaan, T. F. Dann, and P. F. Brunet, "Air pollution monitoring in Amman, Jordan," Journal of the Air & Waste Management Association, vol. 42, no. 6, pp. 814–816, 1992.

[9]. M. Yang and J. A. de Loera, "A Machine Learning Approach to Evaluate Beijing Air Quality," 2018.

[10]. K. Veljanovskal and A. Dimoski, "Air quality index prediction using simple machine learning algorithms," International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), 2018.

[11]. H. Peng, "Air quality prediction by machine learning methods", University of British Columbia, 2015. DOI: 10.14288/1.0166787

[12]. D. Zhu, C. Cai, T. Yang, and X. Zhou, "A machine learning approach for air quality prediction: Model regularization and optimization," Big data and cognitive computing, vol. 2, no. 1, p. 5, 2018.

[13]. P.-W. Soh, J.-W. Chang, and J.-W. Huang, "Adaptive deep learning-based air quality prediction model using the most relevant spatial-temporal relations," IEEE Access, vol. 6, pp. 38186–38199, 2018.

[14]. H. Wang, C. Tseng, and T. Hsieh, "Developing an indoor air quality index system based on the health risk assessment," Proceedings of indoor air, 2008.

[15]. A. Liaw, M. Wiener, and others, "Classification and regression by random Forest," R news, vol. 2, no. 3, pp. 18–22, 2002.

[16]. D. G. Kleinbaum, K. Dietz, M. Gail, M. Klein, and M. Klein, Logistic regression. Springer, 2002.

[17]. K. Nahar, A. Jaradat, M. Atoum, and F. Ibrahim, "Sentiment analysis and classification of arab jordanian facebook comments for jordanian telecom companies using lexicon-based approach and machine learning," Jordanian J. Comput. Inf. Technol., vol. 6, no. 03, pp. 247–263, 2020.

[18]. K. NAHAR, R. Khatib, M. Shannaq, and M. Barhoush, "An efficient holy quran recitation recognizer based on svm learning model," Jordanian J. Comput. Inf. Technol., vol. 6, no. 04, pp. 392–414, 2020.

[19]. "http://www.moenv.gov.jo/EN/List/Daily_Rates_OF_Air_Pollutants", 2019. [Last Accessed on May 16, 2020]