International Journal of Applied Mathematics

Volume 38 No. 5, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

USA AIR QUALITY: FORECASTING PM2.5 CONCENTRATIONS USING MACHINE LEARNING MODELS

Abeer Nasser Abdullah Alnaamani, Ferddie Quiroz Canlas, Sarachandran Nair

[1]Department of Computing
Muscat College
Muscat, Oman
abeer.alnaamani@gmail.com

[2]Department of Computing

Muscat College

Muscat, Oman

ferddie@muscatcollege.edu.om

[3]Department of Computing

Muscat College

Muscat, Oman

saran@muscatcollege.edu.om

Abstract - Emissions have been a serious topic worldwide and every authority in each country needs to set regulations and penalties to control emissions. These gas emissions have an impact on many sectors like society's health, countries' economies, and ecosystems. For that, it is always needed to build any new or updated regulation to be based on strong evidence and that is machine learning comes to fulfil the need. Machine learning algorithms can read big numbers of records and train them then build prediction models. This step allows the entry of historic data, and the algorithm will read and recognize the pattern of the PM2.5 concentrations. As a result of the prediction models, the forecasting will take a part and predict future values of PM2.5 concentrations in the provided meteorological condition available in the dataset. Using the dataset from Kaggle, the steps started with understanding the dataset and handling the errors that appeared like outlier removal, missing values imputation and eliminating the duplicated records. Then four regression machine learning models were applied which are: linear regression, random forest, neural network, and polynomial regression. These four models have been tested to predict the PM2.5 concentrations and tuned the hyperparameters to attain the best accuracy score measured by the value of R-squared and mean squared error MSE. As a result, the four models achieved a good score where linear regression R-squared value= 75.6%, Random Forest R-squared value= 74.9%, neural network R-squared value=75.4% and Polynomial regression accomplished the best model accuracy with R-squared= 100% and an MSE=0%. Taking one further step after defining the best-fitted model, is to forecast future

Received: August 08,2025

International Journal of Applied Mathematics

Volume 38 No. 5, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

values of the concentrations of PM2.5. Using forecaster code in python and a polynomial regression model to produce forecasting values by reading 500 backward steps of data and then generating the forecasted values.

Keywords—*PM2.5 particular matter of size 2.5 mm, emissions, concentrations, machine learning, regression models.*

I. INTRODUCTION

Around the world, particulate matter (PM) is a very common and harmful air pollutant. In particular, combustion from vehicles, industrial processes, and wildfires causes particulate matter pollution. Particulate matter concentrations and composition vary widely across the globe, yet it has a significant impact on public health in the majority of countries [1]. For that, countries across the world had to set regulations and standards to minimize emissions and keep the quality of air at it is best. Many international agreements are signed as well as local agreements setting the limits of gas emissions that form particular matters. Monitoring the emissions is very critical to keep the air clean which reflects on human health and the environment in general. With the increase of industrial processing and vehicle movement and their impact on climate change, a need has a raised to plan for the future. Prediction of PM2.5 concentrations or even the gases emissions helps to regulate the industrial sector to preserve the environment from pollution and the well-being of humans.

There are many algorithms of machine learning used worldwide for prediction. Once an algorithm is found to fit accurately for a certain data set of a certain area it can generate forecasting for the upcoming decade. Many studies have been conducted in this field proving different machine learning models predicting precisely based on the data size and type of attributes included. There is no one model to win and apply for all, it varies based on the inputs and type of analysis. Hence, when searching on the web about the prediction of PM2.5 concentrations using machine learning algorithms, several studies, articles, blogs and conference papers are available discussing different methods of analyzing numeric data or satellite data, as this issue is having a wide global consideration and aiming to reduce gas emissions to keep the earth clean and sustain for all live beings .

This study used two regression models that were later considered old-fashioned compared to the new machine-learning algorithms by many data scientists. However, later in this paper, it will show how the statistical-based algorithm works successfully when it comes to prediction and forecasting future values to help achieve a global goal to minimize pollution on earth.

This study aims to predict the future values of PM2.5 concentrations for the two instruments: Federal Equivalent Method (FEM) and Federal Reference Method (FRM) based on time series data dated from 2016 until 2020 along with environmental attributes that describe the elevation and temperature and other factors. Besides that, another purpose of the study is to test multi-machine learning algorithms and look for the best model to meet the main aim and compare regression models' performance and new machine learning algorithms.

The authors would like to acknowledge Muscat College for the funding of this research.

Received: August 08,2025

International Journal of Applied Mathematics
Volume 38 No. 5, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Furthermore, four machine learning algorithms have been applied, linear regression, random forest, artificial neural network and polynomial regression. All of the algorithms were conducted using the python language program in Anaconda. The four mechanisms have been used in the same rhythm to be comparable. After data cleaning, the four mechanisms splits the data into train and test data by different splits to optimize the maximum accuracy level. However, Linear regression, random forest and polynomial regression performed well at the beginning but not the neural network, Hence, there was a necessity to finetune the hyper-parameters of the model trying to score a better performance and that was applied to the rest of the algorithms that scored well at first. After finetuning and applying the regularization, linear regression, random forest and polynomial regression models performed very well to predict the PM2.5 concentrations for both instruments, but one of them only scored the maximum model efficiency.

II. Review of Related Literature

Particles in the atmosphere, one of the major air pollutants, are groups of solid or liquid particles suspended in the air that come from various sources and in a variety of shapes and sizes. Furthermore, the majority of particulate matter is produced in the atmosphere's lowest layer. However, the finer the particle size, the deeper it can penetrate the respiratory system, in which adsorption is more efficient [2]. Numerous studies have been conducted in this field in most countries as predicting the particular matter guide the leaders to regulate in terms of keeping air quality at it is best. Despite that many international committees' efforts to improve the air quality among the industrial areas reflect on the neighbouring areas or even countries. As explained, the smaller the particular matter the fastest it moves and transfers in the air.

Techniques for predicting air pollution concentrations fall into two broad categories: simulation-based techniques and data mining-based techniques. a simulation-based approach combines physical for generating meteorological and background parameters- and chemical models. However, this technique suffers from numerical model uncertainties, and the parameterization of aerosol emissions is limited due to a lack of data. Where the data mining-based approach uses statistical or machine learning techniques to detect patterns in time series data between predictors and dependent variables.

A. Study in Iran

The capital of Iran is one of the most polluted cities, so as a focused area to study, Hamed and colleagues[3] used data from Tehran Air Quality Control Company, an hourly record of PM2.5 concentrations. The data used along with other environmental factors referenced as meteorological parameters as Tehran has a variety of weather differences based on its geographical form. The team split the data into 60% for training,20% for testing and 20% for validation. The study used Multiple additive regression trees (MART), a classification and regression tree extension and improvement that uses stochastic gradient boosting (SGB) to transform a sequence of weak learners into a complex predictor. As a result, at each iteration (M), a randomly selected subsample of the training set is chosen, and a tree partitions the pseudo residuals (ỹ) into J disjoint regions RJM. Thus, the final model is made up of hundreds to

Received: August 08,2025

International Journal of Applied Mathematics

Volume 38 No. 5, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

thousands of small trees, each of which has contributed to the overall model's improvement. Recurrent Neural Networks (RNN) are a type of DNN that allows information to persist by using cyclic connections (a loop allows information to be passed from one step of the network to the next). Hence, they specialize in time series processing. However, as the time sequence lengthens, the network forgets to train primary inputs, this problem is called "exploding" or "vanishing" gradients, and it is caused by the RNN architecture. To overcome this issue by using a class of RNN called the long short-term memory (LSTM) model.

The authors also evaluated the proposed models by checking the closeness of the predicted value to the observed value which is called mean absolute error (MAE) over the test data and the coefficient of determination (R2) has been used to know the prediction strength for each model. In conclusion, the results showed that the LSTM model outperformed the other models by having the lowest RMSE and MAE values combined with the highest R2 (0.8) value across the entire study area. Furthermore, based on PM2.5 concentrations, the LSTM model accurately predicted 75% of the air pollution levels, demonstrating the importance of sequential feeding in time series modelling. The authors suggest using a few more models would probably improve the model prediction accuracy, for example multiple linear regression or random forest algorithm to achieve very high score.

B. Study in the USA

During last decade, numerous methods have been presented to predict PM2.5, Like satellite-derived aerosol optical depth, land-use variables, and several meteorological variables serving as significant predictor variables. To estimate daily PM2.5 at a resolution of 1 km * 1 km across the United States, Qian and colleagues [4] used a prediction method that integrated multiple machine learning algorithms and predictor variables. In their study, they used three machine learning algorithms which are Random forest, neural network and gradient boosting where in all three methods multiple predictor variables have been used around 100 predictor variables, for example, satellite data, elevation and meteorological variables. They modelled the three algorithms separately to predict PM2.5 and then they used the geographically weighted generalized additive model to obtain the overall PM2.5 predictions. Cross-validation was conducted for the three algorithms based on region, and seasons, at each monitoring site and overall. The R-squared of the three models separately is 0.855 for the neural network, 0.854 for Random forest, 0.818 for gradient boosting and 0.89 for the additive model. As a result, the good model score performed was because of the 100 predictors used to predict PM2.5 concentrations and they develop a way to treat the missing values

C. Study in Ecuador

At the capital of Ecuador, Quito is facing an increase in a particular matter that exceeds the WHO limit and their national limits as well. Jan and his colleagues [5] applied a spatial visualization of the distribution of fine particulate matter trends in two Quito locations based on the wind (speed and direction) and precipitation parameters. The applied boosted Trees and Linear Support Vector Machines are used to classify different levels of PM2.5, as well as Neural

Received: August 08,2025

International Journal of Applied Mathematics

Volume 38 No. 5, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Network regression and a time series analysis is used to provide insight into the parametric boundaries where the classification models perform properly.

However, this model is based on three fundamental meteorological parameters which are: wind speed, wind direction, and precipitation, all of which have a direct impact on pollution. Two methods have been conducted to predict PM2.5: classification and regression. Results indicate significant reliability in the classification of low (<10 μg/m3) versus moderate (10–25 μg/m3) PM2.5 concentrations for the study area. For the whole study period, there is a strong positive correlation between the actual concentrations and the projected concentrations. The model can estimate PM2.5 concentrations more accurately under more severe weather conditions, as evidenced by the somewhat greater correlation during the rainy season. Thus, the authors argue that because the model has a high prediction efficiency for a city with such a complex topography, it could be applied effectively in all other tropical locations.

D. Study in China

To protect the population from the harmful effects of air pollution in advance, it is essential to precisely anticipate PM2.5 concentrations. The variance in PM2.5 is influenced by a number of variables, including the weather and the number of other pollutants present in urban areas. Bekkar and his colleagues [6] did a study on Beijing data of PM 2.5 concentration using hourly basis data. By merging historical data on pollutants, meteorological information, and PM2.5 concentrations in the nearby stations, they constructed a deep learning method to predict the hourly forecast of PM2.5 concentrations in Beijing, China, based on CNN-LSTM, with a spatial-temporal feature. They looked at how different Deep Learning algorithms, including LSTM, Bi-LSTM, GRU, Bi-GRU, CNN, and a hybrid CNN-LSTM model, performed.

The paper's main findings are that this model has high accuracy and stability and can efficiently extract the temporal and spatial features of the data using CNN and LSTM.

E. Study in Australia

In this study, Sinnot and Guan [7] investigated the use of machine learning methods for PM2.5 pollution prediction. Because of the expanding urbanization of the population and its impact on traffic volumes, pollution event prediction is becoming more and more crucial in big cities. To use the data for machine learning algorithms, it had to be collected and cleaned from a range of heterogeneous sources. They investigated LSTMs, ANN models, and linear regression models. It was discovered that LSTM performed best and could reasonably forecast high PM2.5 concentrations. Although ANN and linear models have limitations in predicting high PM2.5 values, overall performance was respectable. The models that were created and investigated in this paper weren't flawless. There are several possibilities for improving them. First off, different machine learning techniques may outperform the sequence models they have explored here. Many of these differ significantly from linear models in terms of their underlying assumptions. As an illustration, ensemble machine learning techniques like Gradient Boosting Decision Tree may be investigated. Additionally, it has been demonstrated that some machine learning techniques may effectively handle datasets that are unbalanced.

Received: August 08,2025

International Journal of Applied Mathematics

Volume 38 No. 5, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

These techniques compromise on overall accuracy to enhance the performance of specific dataset's subsets. For instance, weightings toward various samples and strategies aimed at data sets themselves, such as potential down sampling approaches, are significant model training techniques.

F. Study in India

The necessity of pollutant level forecasting has grown because of pollution's harmful effects on human health and the environment. Using publicly accessible data for New Delhi, Bhattacharya and Shahnawaz [8] used SVR to anticipate levels of pollutants such NO2, SO2, PM2.5 and PM10 as well as the Air Quality Index (AQI). Results were superior when all variables were included rather than only those chosen via principal component analysis. The model has a 93.4 percent accuracy in predicting the amounts of several pollutants, such as sulfur dioxide, carbon monoxide, nitrogen dioxide, particulate matter 2.5, and ground-level ozone, as well as the Air Quality Index (AQI).

Numerous research papers have been written on AQI prediction. To determine the best reliable machine learning model for predicting air quality, experimentation is also performed. Four machine learning algorithms were trained with air quality data during the testing phase to develop predictive models for forecasting AQI. A literature review is used to choose algorithms like Logistic Regression, Ridge Regression, LASSO Regression, and SVR. Following testing and algorithm training using the Air Quality Data in India, Ridge regression was shown to have the best performance in predicting the AQI. It has the lowest MAE and RMSE and the greatest R-square [9].

G. Study in Indonesia

Urban regions are very active and have a high population density. In decades, the urban population has risen by billions. By 2050, three billion more people will live in cities. Indonesia is one of the nations with the quickest rate of urbanization. Hamami and Dahlan [10] used Dataset which was taken from Jakarta's open data for 12 months with several attributes including gas concentration. This study suggests categorizing air quality using methods for classification such Logistic Regression, KKN, Decision Tree, and Random Forest algorithm. The analysis' findings show that by adjusting the decision tree model's two hyper parameters, the maximum number of leaf nodes and the minimum number of samples split, it is possible to increase accuracy to 100%.

III. METHODOLOGY

A. Data Set Description

The data has been obtained from Kaggle [11]. It is a daily record of PM2.5 concentrations from the U.S. Environmental Protection Agency (EPA) National PM2.5 monitoring program. Records are collected from two instruments: the Federal Equivalent Method (FEM) and the Federal Reference Method (FRM) which are operated simultaneously to record PM2.5 concentrations along with several weather conditions of the day recorded such as temperature,

Received: August 08,2025

International Journal of Applied Mathematics

Volume 38 No. 5, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

wind speed, snow depth, daily sunshine in minutes and many others . The total attributes are 32 and the total entries are 116713 for the years 2016-2020.

There are 32 attributes as shown in Table I along with the description of each attribute and what does it refer to as well as the data type if it was categorical or numeric.

TABLE I. THE ATTRIBUTES DESCRIPTION

	Attributes	Description	Data type
1	Aqs_id	The code identify sampling location	string
2	Date-local	Date in format : MM/DD/YYYY	string
3	Latitude _x	Defines the location in degrees	Numeric/ float
4	Longitude _x	Defines the location in degrees	Numeric/ float
5	elevation	Describe the location elevation in meter	Numeric/ float
6	Dominant_source	The primary source of measured pollution	categorical
7	Measurement_scale	Geographic scope of air quality that monitor captured.	categorical
8	Monitoring_object	Describes the reason for measuring air quality.	categorical
9	tavg	The average temperature in C	Numeric/ float
10	tmin	The minimum temperature in C	Numeric/ float
11	tmax	The maximum temperature in C	Numeric/ float
12	prcp	Daily precipitation in mm	Numeric/ float
13	snow	The snow depth in mm	Numeric/ float
14	wdir	The average wind direction in degree	Numeric/ float
15	Wspd	The average wind speed in Km/h	Numeric/ float
16	wpgt	The peak wind gust in Km/h	Numeric/ float
17	pres	The average sea level air pressure in hPa	Numeric/ float
18	tsun	The daily sunshine total in minutes	Numeric/ float
19	Arithmetic_mean_a	FRM concentration (ug/m3)	Numeric/ float
20	Arithmetic_mean_b	FEM concentration (ug/m3)	Numeric/ float

Received: August 08,2025

International Journal of Applied Mathematics

Volume 38 No. 5, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

21	Poc_a	FRM occurrence code	Numeric/ integer
22	Poc_b	FEM occurrence code	Numeric/ integer
23	Meth_type_a	Refers to method a : FRM	categorical
24	Meth_type_b	Refers to method a : FEM	categorical
25	Method_code_a	The FRM three digit digit monitor descriptor	Numeric/ integer
26	Method_code_b	The FEM three digit digit monitor descriptor	Numeric/ integer
27	Monitor_type_a	The FRM monitors organizational classification	categorical
28	Monitor_type_b	The FEM monitors organizational classification	categorical
29	Probe_height_a	The height of the FRM's sampling probe from the ground in meters.	Numeric/ float
30	Probe_height_b	The height of the FEM's sampling probe from the ground in meters.	Numeric/ float
31	pl_probe_location_a	The location of the FRM sampling probe.	categorical
32	pl_probe_location_b	The location of the FEM sampling probe.	categorical

B. Data cleaning

To clean the dataset, usually it needs to model several parts of this data, such as schema, patterns, probability distribution, and other metadata. When a portion of the data does not match the metadata, the error detection stage identifies this portion as having problems. The error detection stage can reveal a variety of mistakes, including outliers, violations, and duplicates. Finally, the error repair stage generates data changes that are applied to the raw dataset based on the faults found and the metadata that causes those errors [12].

As a start, the data set must be sorted ascendingly by date and then transform the data type of date from string to timestamp numeric to use it as one of the predictors.

1) Removing outlier. An observation is considered an outlier if it differs from other observations in such a way that it raises the possibility that it was produced by a separate mechanism. Moreover, identifying what constitutes a regular data pattern and what constitutes an outlier can be challenging because different data and applications have varied definitions of

Received: August 08,2025

International Journal of Applied Mathematics

Volume 38 No. 5, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

what is normal. Additionally, many outlier detection techniques become less effective as the dataset's dimensions expand.

Using the interquartile range method, that is calculated as the difference between the 75th and the 25th percentile of the data. Using python to remove the outlier, a loop is performed to go through all the attributes to remove the upper or lower outlier.

However, after applying this treatment to all variables, one attribute had to be deleted because it has so many outliers that will impact the testing process later which is 'prcp' the daily precipitation.

2) Duplication. In the dataset used for this study, multiple PM2.5 concentrations have been recorded for the same date. Thus, for better performance and cleaned data, duplication has been removed using python. The original data set has 116711 records, now after removing duplication it becomes 1827. But before conducting the deduplication the data should be sorted ascendingly.

*3) Missing values. I*n this study, random imputation has been conducted to treat most of the missing values in the numeric attributes. After removing duplication some attributes do not have missing values while others don't have much of a record as well. However, the numeric attributes that will be imputed are 'tavg', 'tmin, tmax', 'prcp', 'wspd', 'pres', 'probe_height_a', 'probe_height_b'. The attributes that will be deleted because of missing values are 'snow', 'wdir', 'wpgt', 'tsun'.

C. Data Transforms

The type or distribution of data variables can be changed using data transformations. These methods fall within a broad category and can be used with input and output variables as well [13].

*1) One hot encoding. F*or categorical variables, we need to encode them to a binary number, which is called one hot transform.

*2) Normalization.*Because of the way real-valued numeric variables are stored in computers, there is significantly more variation in the range 0–1 than in the wider range of the data type. As a result, normalization of the scaling of variables to this range might be preferable. Normalization is also required for some algorithms to model the data correctly.

Based on the various range difference in the dataset used, like the concentration attributes the range is between 0.1 – 79 whereas the height of the sampling probe is between 1 – 15 and elevation attributes are between 0 1-1920.

This could cause problems when attempting to combine the values as features during modelling and for that normalization was applied.

3) Features Selection. Big data frequently have high dimensions, which makes it difficult to analyse the data and make decisions. The use of feature selection to analyse high-dimensional data and improve learning efficiency has been demonstrated to be beneficial in both theory and practice [14].

Received: August 08,2025

International Journal of Applied Mathematics

Volume 38 No. 5, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

it is necessary to filter out the irrelevant and redundant features by choosing a suitable subset of relevant features to avoid over-fitting [15]. Because we are seeking the best fit model, there are three ways to select the attributes to keep in the model:

- Filter method: utilizing as a pre-processing step. Any machine learning methods have no influence on the feature selection process. Instead, features are picked based on how well they perform in various statistical tests to determine how strongly they correlate with the response variables.

- Embedded method: combining the advantages of the filter and wrapper approaches. It is put into practice by algorithms with built-in feature selection techniques. Regression techniques like LASSO and RIDGE, which feature built-in penalization algorithms to reduce overfitting, are some of the most well-known instances of these techniques.

- Wrapper method: attempt to use a subset of features to train a model with. choosing to add or remove features from certain subsets based on the conclusions that derive from the prior model. Essentially, the issue is simplified to a search issue. These techniques are typically highly expensive to compute [16].

Filter method or ranking based on correlation is used for this study to filter the features to exclude the features affecting the model.

D. *Modelling*

The four prediction regression models used are linear regression, random forest, neural network and polynomial regression. In these four models, the target variables are PM2.5 concentration of method (a) the FRM and method (b) the FEM.

IV. Results and Discussions

A. *Models Comparison*

To attain the best model performance, the data must go into various steps of treatment. Although it was explained in Data cleaning section above all steps of data cleaning process, there were two steps that have major impacts on the model performance which are outlier removing and deduplication.

In the testing stage, outlier removal was conducted only on the response variables: 'arithmetic_mean_a' and 'arithmetic_mean_b'. The R-squared results of the models were very good especially for artificial neural network but not the mean squared error, R-squared was 0.7 while MSE scored 4.1 same results for linear regression and random forest. The contradiction between the R-squared and MSE value is very high, and that goes back to the limitation of the MSE, any extreme value in the data set will have an impact on the model performance and that is the case here when removing the outlier of the response variables only and not for all attributes.

For the second step, deduplication, when running the models, it takes forever to get the results as the file has more than 50,000 records after removing outliers. However, linear regression

Received: August 08,2025

International Journal of Applied Mathematics

Volume 38 No. 5, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

and random forest were scoring good, but the neural network is scoring negative R-squared. Moreover, polynomial regression scored 80% and MSE 0.7% which is quite a good score but not like applying deduplication which leads to 100% of data fitted.

Below, is a detailed comparison of the performance between the four models for both methods (a) and (b). It has been chosen to include the best-fitted model among the variety of hyperparameters for each model, so the competition will be between the best of the best for each model.

1) Method (a)

In Table II, the best results of each of the four models for method (a). To compare between them at different data splits, but with the best hyper parameter. Polynomial regression at degree=1 performed the best fitted model to predict PM2.5 concentrations for method (a) and it can explain 100% of the fitted data with a mean squared error=0% among the other models. Second best model is linear regression; it could explain 75.6% of the fitted data to predict the PM2.5 concentrations when using method(a). In Table III it shows that the result is supported by mean squared error=0.96% as the difference between the actual and predicted value is almost 1%. Third best model is artificial neural network which is scoring very similar to linear regression, it can explain 75.4% of fitted data at data split 70%-30% to predict the response variable with a difference between response and dependent variable is close to 1%.

Lastly, random forest had a very similar performance to linear regression and neural network at different data split. The R-squared of the model scored 74.9% at it is best and the minimum difference between actual and predicted PM2.5 concentrations was=0.99% as stated in Table III.

At the end, to achieve the best prediction for PM2.5 values is to use the best model polynomial regression as it can predict the response variable very accurately based on all the meteorological variables included.

Below are Tables II and III with results of R-squared and MSE for each model to compare and highlight the best model.

TABLE II. R2 VALUES OF ALL MODELS FOR METHOD(A)

Evaluation		**R^2**		
	Model/ splits	**80-20**	**70-30**	**60-40**
Linear Regression	Lasso/L2 at alpha =0.0001	75.6%	75.0%	73.3%
Random Forest	Max_depth= 3	74.9%	74.8%	73.3%
Artificial Neural Network	lbfgs/ layer(10,3)	74.7%	75.4%	72.8%

Received: August 08,2025

International Journal of Applied Mathematics

Volume 38 No. 5, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Polynomial Regression	degree=1	100%	100%	100%

TABLE III. MSE VALUES OF ALL MODELS FOR METHOD(A)

Evaluation		**MSE**		
	Model/ splits	**80-20**	**70-30**	**60-40**
Linear Regression	Lasso/L2 at alpha=0.0001	0. 96%	1.04%	1.11%
Random Forest	Max_depth= 3	0. 99%	1.05%	1.11%
Artificial Neural Network	lbfgs/ layer(10,3)	0.99%	1.02%	1.14%
Polynomial Regression	degree=1	0%	0%	0%

Figure 1 translates Table II and Table III into graphs. The four models at their best performance are shown with the comparisons between the predicted value and actual value. Polynomial regression is the best as the actual and predicted values are equal while the others vary.

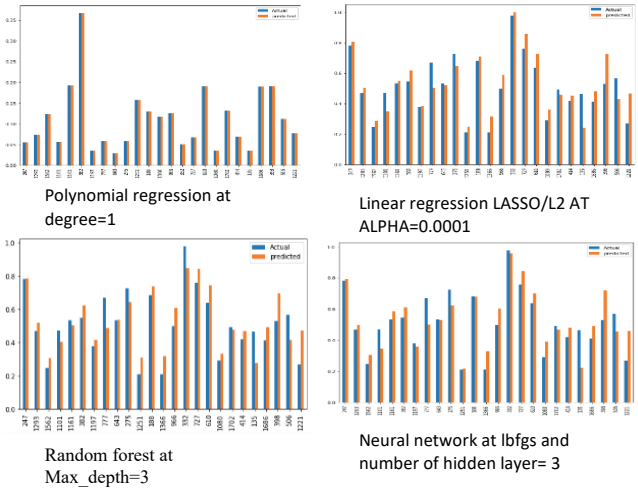

Fig. 1. Bar graph of the four model for method (a)

2) Method (b)

For method (b) as shown in Table IV and Table V, the results of the four models are almost like the method (a), the best-fitted model is polynomial regression that can explain the fitted data with 100% accuracy and as result in the mean squared error =0%. The second best fitted model is linear regression which can predict 75.3% of the predicted value with a 0.86% gap between the actual and predicted PM2.5 concentrations values. Linear regression model result is very similar to the random forest performance that can explain 75% of the fitted data and the

Received: August 08,2025

International Journal of Applied Mathematics

Volume 38 No. 5, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

difference between the actual and predicted concentrations value is 0.87%. Finally, the artificial neural network scored R-squared= 74.6% and the best mean squared error achieved is 0.89%. Likewise, polynomial regression performed the best to predict PM2.5 concentrations based on all meteorological variables.

TABLE IV. R2 VALUES OF ALL MODELS FOR METHOD(B)

Evaluation		**R^2**		
	Model/ splits	**80-20**	**70-30**	**60-40**
Linear Regression	Lasso/L2 at alpha=0.0001	75.3%	74.5%	73.1%
Random Forest	Max_depth= 3	75.0%	73.5%	72.8%
Artificial Neural Network	lbfgs/ layer(2)	74.6%	73.9%	72.2%
Polynomial Regression	degree=1	100%	100%	100%

TABLE V. MSE VALUES OF ALL MODELS FOR METHOD(B)

Evaluation		**MSE**		
	Model/ splits	**80-20**	**70-30**	**60-40**
Linear Regression	Lasso/L2 at alpha=0.0001	0.86%	0.96%	0.99%
Random Forest	Max_depth= 3	0.87%	1.00%	1.00%
Artificial Neural Network	lbfgs/ layer(10,3)	0.89%	3.79%	1.02%
Polynomial Regression	degree=1	0%	0%	0%

In the Fig2. below, it shows similar results as in method (a). The four models at their best performance are shown with the comprison between the predicted value and actual value. Again polynomial regression is the best as the actual and predicted values are equal while the others varies.

Received: August 08,2025

International Journal of Applied Mathematics

Volume 38 No. 5, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

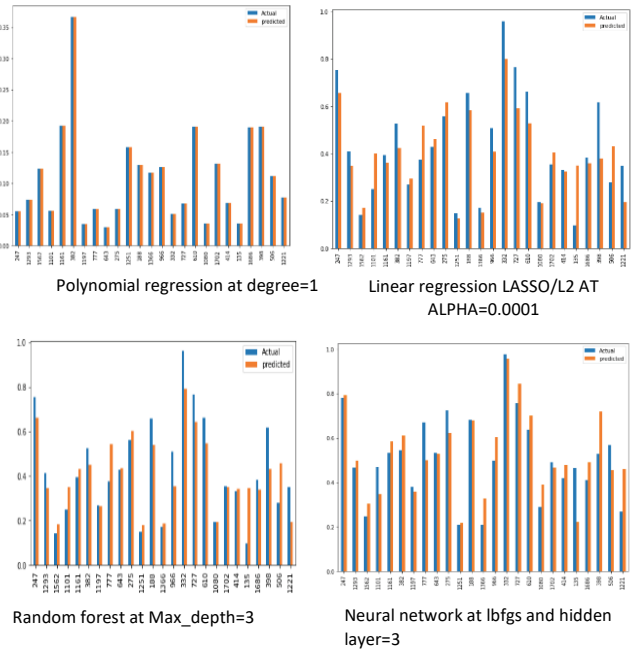

Fig. 2. Bar graph of the four models for method (b)

B. Best fitted forecasting model

As for both method (a) and method (b) the best-fitted model that explains the data set to predict PM2.5 concentrations is polynomial regression. To meet the aim of the study which is to forecast future values of PM2.5 concentrations here is a try to forecast using the ready code of forecaster that was conducted in the best-fitted model polynomial regression. Using Forecaster in python to fit the response train data or the test data part for the polynomial regression the results were very alike as it is a best-fitted model. Figure 3 shows a capture of the forecasted data based on the last 500 records and 200 records back, it will forecast future values and result for method (a) as follows:

```
#prediction based on last 500 record
predictions = forecaster.predict(steps=500)
print(predictions.head(50))
```

```
1461    0.435680
1462    0.423037
1463    0.443203
1464    0.424921
1465    0.434142
1466    0.431009
1467    0.434474
1468    0.434667
1469    0.433184
1470    0.433237
1471    0.432074
1472    0.432627
1473    0.432228
1474    0.432492
1475    0.432291
1476    0.432286
1477    0.432272
1478    0.432274
1479    0.432325
1480    0.432316
1481    0.432340
1482    0.432327
1483    0.432336
1484    0.432335
1485    0.432338
1486    0.432339
1487    0.432338
1488    0.432338
1489    0.432336
1490    0.432337
```

Fig. 3. apture of forecasted PM2.5 concentration using the training dataset

Received: August 08,2025

International Journal of Applied Mathematics

Volume 38 No. 5, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

```
predictions = forecaster.predict(steps=200)
predictions.head(20)
```

```
366    0.389895
367    0.427371
368    0.406267
369    0.416267
370    0.463235
371    0.403895
372    0.408335
373    0.426593
374    0.430421
375    0.428724
376    0.426649
377    0.432376
378    0.427315
379    0.426811
380    0.430007
381    0.429606
382    0.427849
383    0.427233
384    0.428094
385    0.427778
```

Fig. 4. capture of forecasted PM2.5 concentration using the testing dataset

When using the trained data set which is 80% of the data=1460, the forecasting will start from the 1461 number and continues to produce future values, in the other hand, when using the testing data set which has 20% of the data=365 record, the forecast will start from the record 366. In both scenarios, the forecasting was very similar, and changes slightly differently based on the number of steps which means going back 500 or 200 steps back from the actual value than starting to forecast for future values. Figure 5 indicates that for method (b) is quite similar, the left represents the results of the training dataset with 500 steps back and the right represents the forecast output of the test data set with 200 steps back.

```
1461    0.538241
1462    0.532074
1463    0.550137
1464    0.522027
1465    0.540474
1466    0.532242
1467    0.535720
1468    0.540036
1469    0.536457
1470    0.536884
1471    0.534384
1472    0.534877
1473    0.534158
1474    0.535169
1475    0.534604
1476    0.534796
1477    0.534684
1478    0.534607
1479    0.534683
1480    0.534690
```

```
366    0.524337
367    0.542605
368    0.539558
369    0.521993
370    0.527341
371    0.536307
372    0.528502
373    0.533299
374    0.526832
375    0.526017
376    0.528586
377    0.529022
378    0.528574
379    0.528296
380    0.528722
381    0.529283
382    0.528852
383    0.528884
384    0.528725
385    0.528739
```

Fig. 5. Capture of forecasted PM2.5 concentrations for method (b).

In both methods, the weaker the model performs the weaker the forecasted values will be. When the forecaster applies the random forest at the lowest depth performance or neural network at the weaker solver '*sgd'* the output was constant, meaning that the forecasting values are falling on a straight constant line no matter how many backward steps went to learn to produce the future values. Hence, this also supports that polynomial regression is the best model to predict and forecast the given dataset.

V. CONCLUSION

The main purpose of this paper is to forecast future values of PM2.5 concentrations based on the best model that fits the data. The first step to meeting the goal is by understanding the dataset, what are the components of the data and how well it would encounter the aim. An essential step before applying the models is to make sure of the dataset readiness and the only way to accomplish this important step is by cleaning the data based on its format. The main

Received: August 08,2025

Volume 38 No. 5, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

treatment in the cleaning process that is used is outlier removal, missing values imputation and eliminating the duplicated data per each date. The impact of each treatment was very clear in terms of the model's accuracy results, which it could describe the data cleaning as the main key in big data projects to accomplish the best of each model.

The finding of this study was successfully accomplished because of the different phases that took a place before testing the models effectively the data understanding and description, data cleaning procedures and training of the data. The output was producing three good machine learning models to predict the response variable PM2.5 concentrations and one excellent model that will predict the concentrations perfectly which is polynomial regression. However, the regression models performed better than neural networks and random forests which achieve aside the goal of this study and that may be because regression models can work perfectly with big data or a small number of records while neural networks and random forest work best at the big record of data. When the best of the four machine learning models was found, it was used to achieve the main aim to forecast PM2.5 concentrations more accurately than the other machine learning models.

Evaluation is an important procedure to do in a multitasking project, it allows us to know the strengths and weak points in each step of the research.

- Data understanding: this step took two places to make sure the area to be studied matches the dataset obtained, the research of similar studies conducted in the same field to predict PM2.5 concentrations and in chapter two data description gives an overview of the dataset content with its description.

- Data cleaning: The most important step of all, all treatments used in chapter two showed a great impact on the model accuracy results. Many procedures of data cleaning are used and explained in detail in the section Data had either a major or small impact to improve the four presented machine learning models.

- Model training and testing: When it comes to comparing and choosing the best machine learning model, few steps are applied, and many hyperparameters have tuned to achieve the maximum positive performance. First, data splits to train and test data set in three categories: 80% train data-20% test data, 70%-30% and 60%-40%. Second, each model has been tuned using it is different hyperparameters. In Linear regression, it was tuned using Ridge and Lasso with variety of alpha values and in polynomial regression, the model has been tuned by the model degree, but in this case, it has been only tested at degree=1 due to the machine limitation act. The random forest has been tuned using the maximum depth while the neural network has been tuned at different types of solvers.

- Model evaluation: to evaluate the machine learning models, two assessment methods were used: R-squared and means-squared error (MSE). MSE is supporting R-squared results and can detect if R-squared scored the opposite of the real case because of the limitation if the

Received: August 08,2025

Volume 38 No. 5, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

highly correlated variable is added to the model, R-squared might not be affected as it should and vice versa.

- Forecasting: using the forecaster code to produce the future concentrations of PM2.5 was a good way to represent the outcome when having many predictors variables, it will forecast based on the last 200 of back data and produce an infinite number of forecasting. Maybe this could be a small limitation when it is not possible to produce a specific number of forecasting to generate and require further coding enhancement.

Further steps could be done in future based on this research and that goes back to the rich data set used in this paper:

- Separate the file based on location (altitude and longitude) and predict the PM2.5 concentrations taking in consideration the weather condition difference in each location and visualize the results in a topographic map.
- Comparison between method (a) and method (b) and which method is best to used beside all the meteorological variables.
- Enhance the forecaster code to be able to select the number of future values produced.

REFERENCES

[1] S. Bishop, "Air Quality Measurements Series: Particulate Matter", clarity, 2021.

[2] H. Karimian et al., "Evaluation of Different Machine Learning Approaches to Forecasting PM2.5 Mass Concentrations", *Aerosol and Air Quality Research*, vol. 19, no. 6, pp. 1400-1410, 2019. Available: 10.4209/aaqr.2018.12.0450.

[3] H. Karimian et al., "Evaluation of Different Machine Learning Approaches to Forecasting PM2.5 Mass Concentrations", *Aerosol and Air Quality Research*, vol. 19, no. 6, pp. 1400-1410, 2019. Available: 10.4209/aaqr.2018.12.0450.

[4] Q. Di et al., "An ensemble-based model of PM2.5 concentration across the contiguous United States with high spatiotemporal resolution", *Environment International*, vol. 130, p. 104909, 2019. Available: https://www.sciencedirect.com/science/article/pii/S0160412019300650.

[5] J. Kleine Deters, R. Zalakeviciute, M. Gonzalez and Y. Rybarczyk, "Modeling PM2.5 Urban Pollution Using Machine Learning and Selected Meteorological Parameters", *Journal of Electrical and Computer Engineering*, vol. 2017, pp. 1-14, 2017. Available: https://www.hindawi.com/journals/jece/2017/5106045/.

[6] Bekkar, A., Hssina, B., Douzi, S., & Douzi, K. (2021). Air-pollution prediction in Smart City, deep learning approach. *Journal of Big Data*, *8*(1). https://doi.org/10.1186/s40537-021-00548-1

Received: August 08,2025

Volume 38 No. 5, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[7] R. O. Sinnott and Z. Guan, "Prediction of Air Pollution through Machine Learning Approaches on the Cloud," *2018 IEEE/ACM 5th International Conference on Big Data Computing Applications and Technologies (BDCAT)*, Zurich, Switzerland, 2018, pp. 51-60, doi: 10.1109/BDCAT.2018.00015.

[8] S. Bhattacharya and S. Shahnawaz, “Global survey,” *arXiv.org e-Print archive*. [Online]. Available: https://arxiv.org/ftp/arxiv/papers.

[9] A. C. Gogineni and V. Sri Naga Manikanta Murukonda, Prediction of Air Quality Index Using Supervised Machine Learning, May 2022.

[10] F. Hamami and I. A. Dahlan, “Air Quality Classification in urban environment using Machine Learning Approach,” *IOP Conference Series: Earth and Environmental Science*, vol. 986, no. 1, p. 012004, 2022.

[11] "U.S. Air Quality, PM2.5 Concentrations, 2016-2020", Kaggle.com, 2022. [Online]. Available: https://www.kaggle.com/datasets/brannonseay/us-air-quality-pm25-concentrations-20162020.

[12] I. Ilyas and X. Chu, Data Cleaning. San Rafael: Morgan & Claypool Publishers, 2019.

[13] J. Brownlee, Data Preparation for Machine Learning: Data Cleaning, Feature Selection, and data transforms in python. 2020.

[14] J. Cai, J. Luo, S. Wang and S. Yang, "Feature selection in machine learning: A new perspective", Neurocomputing, vol. 300, pp. 70-79, 2018. Available: 10.1016/j.neucom.2017.11.077.

[15] A. Bommert, X. Sun, B. Bischl, J. Rahnenführer and M. Lang, "Benchmark for filter methods for feature selection in high-dimensional classification data", Computational Statistics & Data Analysis, vol. 143, p. 106839, 2020. Available: 10.1016/j.csda.2019.106839.

[16] "What are Machine Learning Models? - Databricks", Databricks, 2022. [Online]. Available: https://databricks.com/glossary/machine-learning-models#:~:text=A%20machine%20learning%20model%20is,sentences%20or%20combinations%20of%20words.

Received: August 08,2025