*J. Int. Environmental Application & Science*, Vol. 20: 80-87 (2025)
*Research Paper*

# Impact of Programming Language on Air Quality Estimation[#]

Emre Dalkiliç[1, *], Şükrü Dursun[1];

[1,2]*Environmental Engineering Department, Engineering & Natural Sciences Faculty, Konya Technical University, Konya, Türkiye.*

*Received March 13, 2025; Accepted March 19, 2025*

**Abstract:** The world has started to gain extra awareness about human health and environmental health after the coronavirus outbreak. In parallel with the increasing environmental awareness, components such as the use of natural resources and the possibility of causing global environmental problems to have started to play an effective role in decision-making processes rather than the financial side of the projects that come to the agenda. States carry out various environmental policies through their ministries, such as preparing legislation on air quality protection and sources affecting air pollution, odor emissions, determining targets, principles, policies and strategies, determining, implementing and having implemented procedures, principles and criteria for the creation of air pollution maps and the preparation of clean air action plans. However, the current situation is no longer sufficient for policymaking, and it is necessary to foresee the future and take steps in this direction. Being able to see today through the eyes of tomorrow provides great convenience in combating problems before they reach the threshold of a crisis in military, political and economic terms as well as environmental terms. Machine learning, a sub-branch of computer science developed in the early 20th century from digital learning and pattern recognition studies in artificial intelligence, is a system that investigates the operability and writing of algorithms that can learn as a structural function and make predictions on data. Written algorithms are designed to learn, instead of following program instructions to the letter, to create data-based predictions from the inputs provided to the system and to act as a decision maker. In the future, there is a need for algorithms that can be written using programming languages to predict air pollution and to determine its effects on public health. Today, using machine learning methods to predict air pollution has become more popular with data and data processing capabilities, which are among the most invaluable capitals. In this study, studies on predictability of air pollution with programming languages will be presented.

***Keywords:*** *Programming Languages, Air Pollution, Forecast, Algorithm*

## Introduction

Air pollution, known as the cancer of our age and ranked first in the world problems, is a natural problem that affects humanity itself because of its desire to increase its level of welfare and live in a more comfortable environment. Although the unplanned and unprepared focus of the development and industrialization race that accelerated after the industrial revolution and focused solely on economy overshadowed environmental and environmental health problems for a while, air pollution that emerged over time has become a danger that we breathe, visible or invisible, physically, biologically or chemically, and that grows every day and will affect humanity at the cost of its life (Varınca *et al.,* 2008). Air pollution causes changes in air compounds in the atmosphere, climate change, acid rain, structural deterioration in the ozone layer, and the effects on human and wild environments. Factors such as locating residential and industrial areas without considering environmental factors (pressure, temperature, wind, humidity, precipitation, solar radiation, *etc.*) of settlements, inadequate green areas, inadequate pollutant control equipment used in industries, poor quality fuel used in vehicles and heating, and lack of maintenance and incomplete combustion of combustion devices used greatly affect air pollution in regions (Kampa & Castanas, 2008; Toros & Bağış, 2017; Yılmaz *et al.*, 2020). Artificial intelligence has become a frequently mentioned focus of attention in the media in the last few years. It is frequently encountered, especially in technology-focused publications, that it will be an indispensable

*Corresponding: E-Mail: emredalkilic95@gmail.com, Tel: +905051414994;*
[#]*This paper has been presented from Ph.D. Thesis of Emre Dalkılıç*

part of the future and can be adapted and used in every sector. Intelligent virtual assistants, unmanned land, air and sea vehicles, and chatbots are preparing humanity for an imaginary future. The idea of a world where labor will decrease and where we can realistically consult and get realistic ideas before or after starting any work foresees a sometimes frightening and sometimes promising future (Chollet & Chollet, 2021). Scientific research is always a pioneering resource in every subject and carries research forward. Compiling these facilitates our inferences on the subject. The aim of this study is to compile studies on the predictability of air pollution with machine learning methods and thus its place in environmental awareness. The accuracy of data for the future is a development that excites the scientific world.

**Machine Learning**

Intelligent Machines was a term that introduced the world to another area in the 1950s where machines were trying to be intelligent like us humans. This was the first step towards a new era. In 1948, Turing and Champernowne invented 'paper and pencil' chess. It was the world's first chess playing computer program. The first AI program that included learning was called "response learning program" and "shopping program" written by Anthony Oettinger in 1951. This was the main effort to learn machines (Shinde & Shah, 2018).

**Machine Learning Steps**

Vapnik and Cortes' reaching a definite result and robust assumptions with support vector machines in 1995 is remembered as the greatest first success of machine learning (Angra & Ahuja, 2017). Following those years, Freund and Schapire developed strengthened weak classifiers, also known as AdaBoost, in 1997. AdaBoost, in which each node is selected from a random subset of features and each of which is formed by a random subset of examples, was discovered by Breiman in 2001. As we approach the present day, Deep Learning systems began to develop and are currently widely used machine learning algorithms such as Naive Bayes (NB), k-Nearest Neighbor (k-NN), Decision Tree, Support Vector Machines (SVM), Bayesian Network, Random Forest, Linear Classifier, Artificial Neural Network (ANN), Logistic Regression and Bootstrap Aggregation (Bagging) (Singh *et al,* 2016). As of 2017, the products available are: Colaboratory (or "Colab" for short), Google Cloud AutoML, KNIME, TensorFlow, WEKA, Torch/Pytorch, RapidMiner, Azure Machine Learning Studio, Accord.NET, Scikit-Learn, Apache Singa, Shogun, Apache Mahout, Apache Spark MLib. Any machine learning application can be implemented using the above interfaces.

**Air Pollution Forecasting Using Machine Learning Methods**

The decrease in the cost of measuring air pollutants and the increase in environmental data and laboratory analysis data has increased the number of pollutant data sets from the past to the present. These set values have many results and complicated correlations. Traditional epidemiological models are becoming very complex and difficult to analyze and interpret the results of large data sets. As a result, in order to facilitate our understanding of the data, data mining and machine learning methods, which are not new but have an expanding scope of analysis methods, offer methods that perform well on similar problems, provide a wide perspective and are highly reliable (Bellinger *et al*., 2017).

**Machine Learning Algorithms and Data Mining**

The ever-evolving big data algorithm belongs to the prediction model. Choosing the prediction models that are created is a big issue on its own. In order to help those who are new to this subject and in the beginning stage in the application of machine learning algorithms, Domingos also discusses some important issues (Domingos, 2012). When choosing the algorithm, the programmer should consider the data complexity of the current problem, and the number of data sets available. For example, a complex, nonlinear classifier will be ineffective on a simple, linear classification problem. Large amounts of data will facilitate the use of advanced learning algorithms such as deep neural networks, but will also force users to consider questions about storage, memory, and training time (LeCun *et al.*, 2015).

As a result, it is widely understood that there is no magic bullet when it comes to choosing machine learning algorithms. From an implementation perspective, it is good practice to select a small, diverse set of algorithms from the paradigm of related methods, test them individually, and choose the one that best meets the performance goals. Alternatively, grouping several sets of models to create an ensemble of predictions has been shown to be an effective solution in theory and practice (LeCun *et al*., 2015). For example, an ensemble of neural networks, support vector machines, Gaussian processes, decision trees, and random forests were applied in the reviewed literatüre (Lary *et al*., 2014).

As a result of the research, the selected algorithms are tested on the existing data set. The resulting models are evaluated to select the model that gives the closest and lowest deviation to the correct result in the forward-looking prediction task. Data mining is the computational basis for the process of analyzing existing big data to solve a problem, determining the limits, revealing valid data and taking these into account to predict the consequences of future and yet unknown events.

Machine Learning, Artificial Intelligence, Mathematics, Database Systems and Statistics are multidisciplinary computational methods used in the estimation process. In addition to basic computational methods, it may be necessary to apply the data mining process and various preprocessing steps to be able to reach a conclusion. A post-processing stage is typically used to visualize the recognized patterns or the received information in an intuitive and easy-to-communicate way.

The most commonly used learning algorithms in air pollution epidemiology can be categorized as predictive or data mining methods (Köktürk *et al*., 2009). Value estimation is a frequently used area of data mining, as it involves determining variables from data taken from specific sample data sets and estimating different data sets considering them. Depending on the variables and structure of the application area, either a data mining algorithm that makes classifier predictions (k-means, k-medians, *etc*.) or numerical predictions (linear regression, MARS, *etc*.) will be selected.

Typical classification, numerical predictions and many methods are given in Table 1. The most appropriate method to choose the data mining method to be used to solve a problem encountered will be to try many algorithms repeatedly and choose the algorithm that gives the best performance values. In other words, prediction algorithms are generally initiated through a controlled learning process. Therefore, the aim is to make b predictions about a example of the target problem. For this, a parameterized function F: a → b is induced.

**Table 1** Machine Learning Algorithms

| **No** | **Deep Learning** | **Neural Networks** | **Regularization** |
|---|---|---|---|
| 1 | Boltzmann Machine | Perceptron | LASSO |
| 2 | Deep Belief Networks | Back-Propagation | Elastic Net |
| 3 | Convolutional Neural Network | Hopfield Network | Least Angle Regression (LARS) |
| 4 | Stack Auto-Encoders | Ridge Regression | |
| **No** | **Rule System** | **Instance Based** | **Clustearing** |
| 1 | Cubist | k-Nearest Neighbour | k-Means |
| 2 | One Rule (OneR) | LVQ | k-Medians |
| 3 | Zero Rule (ZeroR) | SOM | Expectation Maximization |
| 4 | RIPPER | LWL | Hierarchical Clustering |
| **No** | **Ensemble** | **Regression** | **Bayesian** |
| 1 | Random Forest | Linear Regression | Naive Bayes |
| 2 | GBM | OLSR | AODE |
| 3 | Boosting | Stepwise Regression | Bayesian Belief Networks (BBN) |
| 4 | Boorstrapped Aggregation (Bagging) | MARS | Gaussian Naive Bayes |
| 5 | AdaBoost | LOESS | Bayesian Network (BN) |
| 6 | Stacked Generalizationneralization | Logistic Regression | |
| 7 | GBRT | | |
| **No** | **Decision Tree** | **Dimensionality Reduction** | |
| 1 | Classification and Regression Tree | PCA | |
| 2 | ID3 | PLSR | |
| 3 | C4.5 | Sammon Mapping | |
| 4 | C5.0 | Multidimensional Scaling | |
| 5 | CHAID | Projection Pursuit | |
| 6 | Decision Stump | PCR | |
| 7 | Conditional Decision Trees | PLSDA | |
| 8 | M5 | MDA | |
| 9 | | QDA | |
| 10 | | RDA | |
| 11 | | FDA | |
| 12 | | LDA | |

For the value estimates to be continuous, it can be applied to real numbers such as b ∈ R, but also to integers such as b ∈ I. The most preferred algorithms in supervised learning algorithms are support vector machines, artificial neural networks, Bayesian methods and decision trees. The main reason why decision tree algorithms are shown as an effective classifier is that they have independent variables and rules. It creates a simple estimation method by dividing data sets into sections with the divide and conquer strategy. Nowadays, there are many methodologies and algorithms for machine learning. Machine learning is divided into three groups according to the learning method: Supervised, Unsupervised and Reinforcement.

**Supervised Machine Learning Algorithms**

Supervised machine learning algorithms are the most widely used algorithms among machine learning algorithms. In these model algorithms, the data scientist teaches the algorithm what results to produce in a teacher-like manner. The algorithm is trained with data sets with pre-labeled outputs, like a small child trying to learn the multiplication table by heart.

Supervised machine learning algorithms vary depending on the problem to be addressed. The algorithms we will choose for each problem differ from each other in terms of their features. Naive Bayes classifier, k-nearest neighbor, SVM (Support vector machines), Linear regression, Decision trees, Neural networks are some of the supervised learning algorithms.

**Unsupervised Machine Learning Algorithms**

In unsupervised learning, outputs are not included in the study. Observed units are brought together according to their similar features. Unsupervised learning, which uses machine learning algorithms to interpret and classify unidentified data sets, classifies data without requiring any intervention.

## Result and Discussion

In the article titled Time Series Analysis and Forecast for Air Pollution in Ankara: Box-Jenkins Approach by (Turgut & Temiz, 2015), they aimed to estimate the future values of PM10 pollutant in Ankara using Box-Jenkins methodology. Box-Jenkins methodology works as a single variable model in estimating the future status of a parameter. It is a successful method in near-term estimations. In the study, 232 data were used using 5-year weekly data from Ankara Sıhhiye station and time series were used using Minitab Package program. Autocorrelation graphs were created, and it was observed that the time series were stationary and Extended Dickey-Fuller Unit Root Test, EViews package program was used for proof. According to Dickey-Fuller Unit Root test, since the critical values were greater than the statistical value, $H0$ hypothesis was rejected. In other words, the time series is stationary. By examining the graphs and proof results, it was decided to apply ARIMA (3,0,0) model of Box-Jenkins approach accordingly. When 232 data were processed into the model program and the result output was obtained, it was estimated that the PM10 pollutant in the Ankara Sıhhiye region would be at an average level of 83.21 mg/m3. It was concluded that it was at a medium pollution level.

Özel (2019) titled Air Pollution Prediction for Ankara Province Using Markov Chain, it was aimed to estimate long-term air pollution values using Markov chains by using $PM_{10}$ air quality index values at Bahçeli station in Ankara city center. Markov process is a system that does not need past data other than previous processes. With the help of air quality and daily changes data obtained from the ministry page, the transition matrix giving the probabilities was reached by using the number of these changes. For example, when the air quality is good in Ankara, the probability of the weather being good the next day is 40%, the probability of it being middle class is 40%, the probability of it being sensitive is 17% and the probability of it being unhealthy is 3%. To obtain the limit distribution that will show the long-term structure of the Markov chain, the equilibrium distribution was reached by using the MATLAB program. As a result, the long-term probability of air quality in Ankara being good is 46%, the probability of being moderate is 19%, the probability of being sensitive is 14%, the probability of being unhealthy is 4.5%, the probability of being poor is 2.1% and the probability of being hazardous is 15%.

Kaplan *et al.*, (2014) in their study titled "Prediction of $PM_{10}$ and $SO_2$ substances causing air pollution using artificial neural networks and calculation of error rate", it was aimed to predict $PM_{10}$ and $SO_2$ pollutants causing air pollution using Levenberg-Marquardt learning algorithm. Levenberg-Marquardt algorithm, which is one of the learning algorithms of feedback model in artificial neural networks, is a system that approaches the error surface parabolically at each iteration stage and the minimum parabola angle represents the result for that step. In this study, Wind, Humidity, $PM_{10}$ and $SO_2$ data taken every hour for 5 days belonging to Kütahya province were used. 96 of the data used were used as training and 24 as test data. A 3-layer feedback network structure was created using MATLAB artificial neural networks toolbox. The tangent sigmoid function was used as the activation function, and 30 neurons were selected in the intermediate layer, and the weights were changed in the feedback using the Levenberg-Marquart learning algorithm. As a result, because of using 5-day air pollutant concentration and meteorological data in Kütahya province, the pollutant estimate of the 6th day was realized, and the estimates largely coincided with the real data. According to this model, the normalized PM10 root mean square error value was 0.0161 and the $SO_2$ root mean square error value was 0.0372.

In the study titled "A Comparative Evaluation of Air Pollution Prediction with Machine Learning Algorithms" conducted by (Gültepe, 2019) on developing models to predict air pollution with various machine learning algorithms in Kastamonu province, Artificial Neural Networks (ANN), Random Forest, K-Nearest Neighborhood, Logistic Regression, Decision Tree, Linear Regression and Naive Bayes methods were used as prediction models. In the performance evaluations of the methods used on the Kastamonu DataSet, it was determined that there were statistically significant differences in terms of the Explanatory Coefficient (R2), Mean Squared Error (MSE), Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) metrics. Studies have shown that the artificial neural network model provides 87% success in predictive data accuracy, while Random Forest and Decision Tree algorithms work at 99% predictive data accuracy. The Linear Regression method exhibited a very poor performance with an accuracy rate of 30%.

Study titled "Estimating CO2 emissions in OECD countries with machine learning" studied the estimation of CO2 emissions of 10 OECD countries with machine learning. 80% of the data was used for training and 20% for testing (Garip, 2017). Three methods; M5P, support vector machine (SVM) and artificial neural networks (ANN) were used as machine learning methods. After the estimations were made, statistical functions such as Mean Absolute Error (MAE), Mean Square Error (MSE), Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE) were used to evaluate the performance of the applied machine learning methods. As a result of the estimations, it was seen that the SVM machine learning method made quite successful estimations in 7 out of 10 countries. It was understood that SVM was more successful than ANN and M5P.

Dokuz et al., (2020) provided a study that can prevent selection confusion regarding the problem of which parameters should be examined with which method and how in machine learning, which is used as a tool in improving air quality. In the study titled "Using machine learning methods for the estimation and spatial distribution of air quality parameters", machine learning algorithms were introduced and many studies that performed air pollution estimation with machine learning were examined. As a result of the examinations, it was determined that the data volume to be used in algorithm selection significantly affects the algorithm success; pollutant parameter selection for the application area; the presence of a real air quality station to verify the algorithm result; a good data volume that can take meteorological variability into account; importance should be given to data regularity, quality and precision; The analysis technique should be determined according to the spatial distribution of the estimated concentration amounts obtained, whether it is an urban or rural area, the necessity of defining the topography and the type of land use. Paying attention to these headings will increase the accuracy percentage of the study to be conducted.

Rybarczyk and Zalakeviciute, (2016) in their article titled "Machine learning approach to predict urban pollution" used a model based on machine learning to predict PM2.5 concentrations from air pollutant parameters in Quito, Ecuador, by accepting wind (speed and direction) and precipitation meteorological data as variables. Data collection processes were carried out with PM2.5 and meteorological data measuring devices from two regions determined in Quito. WEKA data mining program was used for machine learning. In addition, modeling studies were conducted with decision tree algorithms. For both regions, PM2.5 concentrations were predicted with a classification based on 15 µg/m3 threshold and 65% accuracy was obtained. Since the complexity in predicting pollution originating from vegetation and meteorological data in Ecuador and the difficulties in producing a model with minimum parameters and maximum accuracy are quite high, 65% accuracy was accepted as a very high accuracy.

Zhan et al., (2017) used a new machine learning algorithm, Geographically Weighted Gradient Boosting Machine (GW-GBM), in their study titled "Spatiotemporal estimation of continuous daily PM2.5 concentrations across China using a spatially explicit machine learning algorithm" in China, where acute human health problems are most frequently encountered in the world. The model used aerosol optical depth and meteorological data as variables to investigate the spatial variation of PM 2.5 concentrations. As a result, it was observed that 95% of the Chinese population lives in areas where PM 2.5 concentration is higher than 35 mg $m^{-3}$.

In the article titled "*Using Ensemble Regression Algorithms to Increase the Prediction Success of Air Quality Index*" by Irmak and Aydilek (2019), the air quality index of Adana province was estimated. The calculations were made with Spyder, a Python programming language compiler that includes useful libraries for data mining and machine learning. They collected hourly measurement values of PM10, $SO_2$, $NO_2$, $O_3$ and CO air pollutants between 2013-2017 from 4 different stations. The regression methods used are random forest regression, decision tree regression, support vector regression, K-EN nearest neighbor regression, linear regression and artificial neural network regression, batch regression, adaptive booster regression, gradient booster regression and sampled total regression. After the calculation process was completed, a data set containing 43838 records for the years 2013-2017 was obtained. A training data set of 75% and a test data set of 25% were created by randomly selecting from this data set. Machine learning algorithms were run on the Spyder compiler with the help of Pandas and Sklearn libraries. Regression algorithms were first trained with the training data set. Then, test data was applied to the same algorithm and error values were recorded. The same operations were applied to each algorithm and their values were recorded. As a result, the algorithm that could best predict the air quality index was random forest regression. Community-based regression algorithms produced more successful results than other algorithms. The fastest algorithm was linear regression with 0.00699 seconds.

In their paper titled "A Machine Learning Approach to Predict Air Quality in California", Castelli *et al*., (2020) used support vector regression (SVR), a popular machine learning method, to predict pollutant and particulate levels and estimate the air quality index (AQI). The Radial Basis Function was the kernel type that allowed SVR to predict with the highest success rate among the alternative algorithms tested. Using all available variables instead of determining features using principal component analysis was a move that increased success in the study. It was observed that SVR with the RBF kernel gave 94.1% success in predicting hourly pollutant concentrations of nitrogen dioxide, carbon monoxide, ground-level ozone, particulate matter 2.5, and sulfur dioxide, as well as the air quality index of California.

They wanted to predict future PM2.5 and PM10 concentrations using daily PM2.5 and PM10 concentration data in Kunming and Yuxi cities of China between 2015 and 2016 using a new hybrid model (CI-FPA-SVM) written by (Li, Kong, & Wu, 2017). At the beginning of the study, it was difficult to obtain results due to the difficulty of understanding the relationship between variables with different definitions. For this reason, they created a hybrid model by combining different models. Here, various parameters were optimized with FPA and a nonlinear system was obtained with SVM. In order to test the hybrid model, six benchmark models were considered, including FPA-SVM, CI-SVM, CI-GA-SVM, CI-PSO-SVM, CI-FPA-NN, and multiple linear regression model. The empirical study results show that the proposed CI-FPA-SVM model is far superior to all the evaluated benchmark models in terms of high prediction accuracy and the application of the model for prediction can lead to more effective air quality monitoring and management.

Sotomayor-Olmedo *et al*., (2013) in their study "Predicting Urban Air Pollution in Mexico City Using Support Vector Machines: Kernel Performance Approach" presents a forecast pollution model using support vector machines and kernel functions such as Gaussian, Polynomial and Spline. In the study where these techniques were used, ozone ($O_3$), particulate matter ($PM_{10}$) and nitrogen dioxide ($NO_2$) in Mexico City were estimated. As a result of the study, it was observed that the Gaussian kernel gave more realistic results for the necessary calculations. It was revealed that the polynomial kernel gave less successful results than the Gaussian.

In a study conducted in 2012 to make short-term estimates of air pollution in Macau, data sets were created by organizing daily meteorological and air pollution data collected from continuous monitoring stations. While SVM had a good coverage ability in the study where Support Vector Machine was preferred in regression and time series estimation, it was observed that the performance of the SVM model generally depends on the kernel selection. In the experiments conducted with different kernel types, the estimation results of the linear model and the RBF model gave results that were quite close to the real data sets for the test of SO2 and NO2 data. Similarly, in the tests of seasonal data, the two models provided a higher accuracy rate than the other three models. However, some delays and underestimations occurred in these two models in the winter experiment. In the studies conducted, it was concluded that using the Linear model and the RBF model in the air pollutant and meteorological data estimations in the city of Macau in five different models yielded good results with lower errors compared to the other data sets. In addition, it was recommended that users who want to make air pollutant estimations in similar cities use these models (Vong *et al*., 2012).

As a result of many factors such as industrialization, uncontrolled urbanization, climate change, violation of regulations by industrial facilities and inadequate inspection activities of supervisory institutions and organizations, the air quality of cities is getting worse over time. Although research shows that air quality is extremely important for human and environmental health, the extent to which this situation will be taken in the future is no longer an unknown through algorithm and modeling programs such as machine learning.

Due to many reasons such as the fact that industrial areas in cities are located within the city center, the city's wind paths are blocked by high-rise buildings, natural gas has not yet become a dominant fuel type in the city, the city administration units such as governorships, mayors, provincial directorates and clean air center directorates are drowning in bureaucratic work and the city's inability to manage air quality well, air quality is above the limit values specified in the regulations in almost all air pollutant parameters, especially in the winter months.

Revealing how this situation will proceed in the coming years is considered as a project that will positively affect the air quality of the city in the future, contribute to the literature and human health. Considering these situations, collecting air quality data and meteorological parameters from previous years covering city centers and districts, processing them and using the appropriate algorithm machine learning algorithm written in the programming tool, estimating the status of pollutant parameters in the coming years will be a project that will take cities quite a bit further in terms of action plans. It is a definite opinion that it would be better if the studies carried out were taken one step further and converted into web or mobile applications, made accessible to the public and operated. In addition, the support of municipalities and ministries in such projects is essential. It is obvious that choosing the right algorithm in machine learning will give the most suitable model for our data. Of course, it is not an easy process to try algorithms one by one for this. Also, doing this manually can leave us with a result such as struggling with the data and eventually getting discouraged.

It is not practical to try everything, so of course it is necessary to use machine learning tool providers AutoML systems. The best ones show themselves with scans on feature engineering algorithms and normalizations. The hyper-parameter setting of the best model or models is left after the algorithm selection. It should be known that this situation is easily solved by AutoML and in my opinion, it is the process that requires the most thought. In summary, machine learning algorithms are only a part of the machine learning puzzle. In addition to algorithm selection (manual or automatic), you need to deal with optimizers, data cleaning, feature selection, feature normalization and (optionally) hyperparameter settings. Once we have considered all of these and created a model that works for our data, it is time to deploy the model and update it as conditions change.

***Acknowledgment**: I would like to thank Environmental Engineer Feride Dalkilic for her contributions to this study.*
***Compliance with Ethical Standards Ethical responsibilities of Authors**: The author has read, understood, and complied as applicable with the statement on "Ethical responsibilities of Authors" as found in the Instructions for Authors".*
***Funding**: There is no funding this study.*
***Conflict of Interest**: The authors declare that they do not have any conflict of interest.*
***Change of Authorship**: The author has read, understood, and complied as applicable with the statement on "Ethical responsibilities of Authors" as found in the Instructions for Authors and is aware that with minor exceptions, no changes can be made to authorship once the paper is submitted.*

## References

Angra S, Ahuja S, (2017) Machine learning and its applications: A review. Paper presented at the *2017 international conference on big data analytics and computational intelligence (ICBDAC)*. https://doi.org/10.1109/ICBDACI.2017.8070809

Bellinger C, Mohomed Jabbar MS, Zaïane O, Osornio-Vargas A, (2017) A systematic review of data mining and machine learning for air pollution epidemiology. *BMC Public Health*, **17**, 1-19. https://doi.org/10.1186/s12889-017-4914-3

Castelli M, Clemente FM, PopovičA, Silva S, Vanneschi L, (2020) *A machine learning approach to predict air quality in California. Complexity,* 2020(1), 8049504. https://doi.org/10.1155/2020/8049504

Chollet F, Chollet F, (2021) Deep learning with Python: Simon and Schuster. https://sourestdeeds.github.io/pdf/Deep%20Learning%20with%20Python.pdf

Dokuz Y, Bozdağ A, Gökçek B, (2020) Hava Kalitesi Parametrelerinin Tahmini ve Mekansal Dağilimi Için Makine Öğrenmesi Yöntemlerinin Kullanilmasi. *Niğde Ömer Halisdemir Ün. Müh. Bil. Der.*, **9**(1), 37-47. https://dergipark.org.tr/tr/pub/ngumuh/issue/52170/654092

Domingos, P. (2012) A few useful things to know about machine learning. *Communications of the ACM,* **55**(10), 78-87. https://dl.acm.org/doi/pdf/10.1145/2347736.2347755

Garip E, (2017) OECD ülkelerindeki $CO_2$ emisyonunun makina öğrenmesi ile tahmin edilmesi. İstanbul *Medeniyet Üni. Fen Bil. Enst. Müh. Tezler.* https://acikbilim.yok.gov.tr/

Gültepe Y, (2019) Makine öğrenmesi algoritmaları ile hava kirliliği tahmini üzerine karşılaştırmalı bir değerlendirme. *Av. Bil. ve Tek. Derg.* (16), 8-15. https://dergipark.org.tr/tr/pub/ejosat/issue/45333/530347

Irmak ME, Aydilek İB, (2019) Hava kalite indeksinin tahmin başarısının artırılması için topluluk regresyon algoritmalarının kullanılması. Academic platform-*J. Engin. & Sci.,* **7**(3), 507-514. https://dergipark.org.tr/tr/pub/apjes/issue/44190/478038

Kampa M, Castanas E, (2008) Human health effects of air pollution. *Environ. Pollut.,* **151**(2), 362-367. https://doi.org/10.1016/j.envpol.2007.06.012

Kaplan Y, Saray U, Azkeskin E, (2014) Hava kirliliğine neden olan PM10 ve SO2 maddesinin yapay sinir ağı kullanılarak tahmininin yapılması ve hata oranının hesaplanması. *AKU J. Sci.Eng.***14** (2014) 025201 (1-6) https://dergipark.org.tr/tr/pub/akufemubid/issue/1611/20175

Köktürk F, Ankarali H, Sümbüloğlu V, (2009) Veri madenciliği yöntemlerine genel bakış. *Turkiye Klinikleri J. Biostat*., **1**(1), 20-25. https://www.turkiyeklinikleri.com/article/en-veri-madenciligi-yontemlerine-genel-bakis-53374.html

Lary DJ, Faruque FS, Malakar N, Moore A, Roscoe B, Adams ZL, Eggelston Y, (2014) Estimating the global abundance of ground level presence of particulate matter ($PM_{2.5}$). Geospatial health, 8(3), S611. https://doi.org/10.4081/gh.2014.292

LeCun Y, Bengio Y, Hinton G, (2015) Deep learning. **Nature**, **521**(7553), 436-444. https://www.nature.com/articles/nature14539

Li W, Kong D, Wu J, (2017) A new hybrid model FPA-SVM considering Cointegration for particular matter concentration forecasting: A case study of Kunming and Yuxi, C*hina. Comput. Intellig. & Neurosci.,* 2017(1), 2843651. https://doi.org/10.1155/2017/2843651

Özel G, (2019) Markov zinciri kullanarak Ankara ili için hava kirliliği tahmini. *Int. J. Sci. & Tech. Res.*, **3**(2), 144-151. https://dergipark.org.tr/tr/pub/bilgesci/article/546317

Rybarczyk Y, Zalakeviciute R, (2016) Machine learning approach to forecasting urban pollution. Paper presented at the *2016 IEEE Ecuador Technical Chapters Meeting* (ETCM). https://doi.org/10.1109/ETCM.2016.7750810

Shinde PP, Shah S, (2018) A review of machine learning and deep learning applications. Paper presented at the 2018 Fourth international conference on computing communication control and automation (ICCUBEA). https://doi.org/10.1109/ICCUBEA.2018.8697857

Singh A, Thakur N, Sharma A, (2016) A review of supervised machine learning algorithms. Paper presented at the 2016 *3rd International Conference on Computing for Sustainable Global Development (INDIACom*). https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7724478

Sotomayor-Olmedo A, Aceves-Fernández MA, Gorrostieta-Hurtado E, Pedraza-Ortega C, Ramos-Arreguín JM, Vargas-Soto JE, (2013) Forecast urban air pollution in Mexico City by using support vector machines: A kernel performance approach. *Int. J. Intel. Sci.* **3**, (3) 126-135 http://dx.doi.org/10.4236/ijis.2013.33014

Toros H, Bağış S, (2017) Hava kirlilik modellerinde kullanılacak emisyon envanteri oluşturulması için yaklaşımlar ve İstanbul hava kirliliği dağılımı örneği. *Çukurova Ün. Müh.-Mim. Fak. Der.*, **32**(2), 1-12. https://dergipark.org.tr/tr/pub/cukurovaummfd/issue/32170/357001

Turgut D, Temiz İ, (2015) Ankara'daki hava kirliliği için zaman serileri analizi ve tahmin: box-jenkins yaklaşımı. *Alphanum. J.*, 3(2), 131-138. https://dergipark.org.tr/tr/pub/alphanumeric/article/20803

Varınca, K. B., Güneş, G., & Ertürk F, (2008) Hava kirleticilerinin insan sağlığı ve iklim değişikliği üzerine etkileri. *U UHAKS,* Konya. https://www.academia.edu/1424169/Hava_Kirleticilerinin

Vong, C.-M., Ip, W.-F., Wong PK, Yang JY, (2012) Short-term prediction of air pollution in Macau using support vector machines. Journal of Control Science and Engineering, 2012(1), 518032. https://doi.org/10.1155/2012%2F518032

Yılmaz M., Toros H., İncecik, Öztürk, Z., Kirkil, G., Öztaş, D, Beba, HE. (2020) Dilovası Hava Kirliliğinin Trafik Emisyonları Açısından Değerlendirilmesi. *Ulusal Çevre Bil. Araş. Der.*, **3**(1), 43-51. https://dergipark.org.tr/tr/pub/ucbad/issue/54360/738990

Zhan Y, Luo Y, Deng X, Chen H, Grieneisen ML, Shen X, Zhang M, (2017) Spatiotemporal prediction of continuous daily PM2. 5 concentrations across China using a spatially explicit machine learning algorithm. *Atmos. Environ.*, **155**, 129-139. http://dx.doi.org/10.1016/j.atmosenv.2017.02.023