International Journal of
Science and Engineering Investigations vol. 6, issue 71, December 2017
ISSN: 2251-8843

Machine Learning Algorithms in Air Quality Index Prediction

Kostandina Veljanovska[1], Angel Dimoski[2]
[1]Department of Intelligent Systems, Faculty of ICT, University "St. Kliment Ohridski", Bitola, Republic of Macedonia
[2]Faculty of ICT, University "St. Kliment Ohridski", Bitola, Republic of Macedonia
([1]kostandinav@gmail.com)

Abstract-Urban air pollution is one of the biggest global problems at this moment. It has effect on human's most significant element of life- oxygen. Air pollution increases risk of diseases and mortality through respiratory and cardiovascular impact, so understanding and predicting of Air Quality Index is significant public health challenge. In this paper we made comparison between three simple machine learning algorithms, neural network, k-nearest neighbor and decision tree. The results are promising and it was proven that these algorithms are very efficient in predicting Air Quality Index.

Keywords- *Machine learning, k-Nearest Neighbor, Decision Tree, Neural Network, Air Quality Index*

I. INTRODUCTION

Today, in time of big and fast technological development, science is aiming on that, artificial intelligence (AI) to become equal to human intelligence. Machine learning is area where system which implements AI gathers data from sensors in an environment and learns how to act. It does that through training similar to the model of human learning. On opposite, traditional programming is making system to work only by orders which are given by programmer, but not to make decision on its own.

This was the reason why we choose machine learning to predict air quality index, wishing to allow the model to adapt to changing circumstances in an environment. Machine learning have two or three phases: training, testing, validation (or training and testing). Biggest part of the data is used to train the system, remaining data is used for other two phases (for example 70% training, 20% test, 10% validation). In continuation of the paper these phases are described and compared through two supervised learning algorithms k-nearest neighbor (k-NN) and Decision Tree (DT) and one unsupervised algorithm Neural Network (NN).

Generalization is ability of machine learning algorithm to give accurate output for new samples which previously were not observed in the training set. During training, algorithm separates key information of training samples, and according to this info it gets picture of their parameters and the logic in the problem domain.

II. MODELING THE PROBLEM

Air pollution in the Republic of Macedonia is concerned as a serious problem since measured values of the parameters of air quality are many times above the limit values determined to protect human health. Situation is very serious in the larger urban areas. This way we made contribution on the scientific level at the first phase in predicting air quality index in order to help improve the situation. In this project we developed three different classifiers based on different algorithms and we use dataset (Table 1) which is based on model of official web site of Ministry of environment and physical planning of Macedonia. Measure stations measure SO_2 (sulfur dioxide), NO_2 (Nitrogen dioxide), O_3 (Ozone), CO (Carbon monoxide), suspended particulates PM2.5 (Fine particles) and PM10 (Large particles).

TABLE I. ATTRIBUTES OF AIR POLLUTION (VALUES ARE EXPRESSED IN $\mu G/M^3$, AND CO IS EXPRESSED IN MG/M^3)

Attributes	Values	Meaning
SO2	1,2,3,4,5	Sulfur dioxide-(0-50,50-100,100-350,350-500,500+)
NO2	1,2,3,4,5	Nitrogen dioxide-(0-50,50-100,100-200,200-400,400+)
O3	1,2,3,4,5	Ozone-(0-60,60-120,120-180,180-240,240+)
CO	1,2,3,4,5	Carbon monoxide-(0-5,5-7.5,7.5-10,10-20,20+)
PM 2.5	1,2,3,4,5	Fine particles-(0-15,15-30,30-55,55-110,110+)
PM 10	1,2,3,4,5	Large particles-(0-25,25-50,50-90,90-180,180+)
AQI	0,1,2	Air Quality Index- (Low, Medium, High)

From Fig.1, it can be seen that border of low/medium/high air pollution is set on index with value 3 (medium). From all six attributes, if two of them are with index 2 and other two are with index 3 then we have medium level of air pollution. If values of PM2.5 and PM10 particles are with index 4 or 5, then again we have high level of air pollution, independently of other attributes. If one of the attributes is with value 3 and others with lower values then 3, then we have low AQI. If attributes are with AQI value 1 or 2 then AQI is low.

For the experiments we use dataset which is constructed according to Air Pollution in Macedonia. Dataset contains 156 samples (13 samples per month of 2016), 33 of this samples are with High Air Pollution Index, other 42 are with Medium Air Pollution Index and the rest 81 samples are with Low Air Pollution Index. Purpose of this project is to build three

classifiers, to train the algorithms with previously measured data and to make these classifiers capable of predicting Air Quality Index with some new measured data. Supervised Test Dataset contains 21 samples and it is used in k-NN and DT algorithms. Neural network works as unsupervised learning using same training dataset of 156 samples with 6 attributes. These samples are classified into 3 different classes: Low, Medium or High.

	A	B	C	D	E	F	G
130	2	1	1	2	3	2	0
131	2	2	1	2	2	3	1
132	3	1	2	3	4	4	2
133	3	1	2	2	2	3	1
134	2	2	2	1	3	2	1
135	1	1	2	1	2	2	0
136	1	2	2	1	3	2	1
137	1	2	3	1	3	3	1
138	1	2	1	2	3	2	1
139	2	3	2	3	4	3	2
140	2	2	1	2	3	4	2
141	3	2	2	3	4	4	2
142	1	2	2	2	5	4	2
143	1	1	1	2	5	5	2
144	1	1	1	1	3	4	2
145	1	2	2	1	2	3	1
146	2	1	1	1	3	2	0
147	2	2	1	1	3	3	1
148	1	2	1	2	1	2	0
149	2	1	1	2	2	2	0

Figure 1. Getting AQI according to values of attributes

Realization of experiments of this project is performed in MATLAB Machine Learning Toolbox like platform for experimenting.

III. Neural Network Algorithm

Neural Network is one of the most popular techniques of machine learning. Created according to structure of biological human neural system, they are built from large number of interconnected neurons, positioned in layers which work together for solving certain problems. These problems can be very complex, but, NN has its own way to solve them. Neural Network can be defined through weights of each of the neuron connections, connection structure of neurons between layers and activation function [1, 2]. It can, also, learn how to solve a problem by using technique called "deep learning" with observing previous samples [3, 4].

Neural Network in this project is classifying samples to "low", "medium" and "high" levels of air pollution, depending on training dataset. In this case there are 6 attributes as input and there are 3 possibilities as output values which need to be predicted. There is hidden layer too, which contains 10 neurons. There are many rules for setting the number of neurons in hidden layer [5]:

(1) Number of neurons in hidden layer need to be half of sum of input and output values (Ni+ No)/2=Nh;

(2) Number of neurons in hidden layer need to be 2/3 of number of input values plus output values Nh= 2/3*Ni+No;

(3) Number of neurons in hidden layer needs to be smaller than half of input value Nh < Ni/2

(4) Number of neurons in hidden layer needs to be half of sum of input and output values plus number in interval from {1, 10}. Nh= ((Ni+No)*0.5)+{1,1}

(5) Number of neurons in hidden layer doesn't have to be too large, because of overcrowding, which reduced accuracy.

A. *Experiments with NN Algorithm*

In our case, number of neurons in hidden layer is 10 (Fig. 2), because experimenting with this value, we get smallest optimal error. In this project the neural network is performed like unsupervised learning, which means that the input data for training are known by network, but output data is unknown, so the network performs classification by knowledge gathered from input data. For this purpose forward propagation is used, which means that, activation of each neurons is determined from previous layer and weights between those neurons. With Forward propagation we get output value and that value is compared with real value to get the error. After the error is known with forward propagation, minimization is done using back propagation. This means that algorithm is propagating backward, from output layer to input layer and on its way it finds error for each of the weights. That value will be changed to minimize the total error [4].

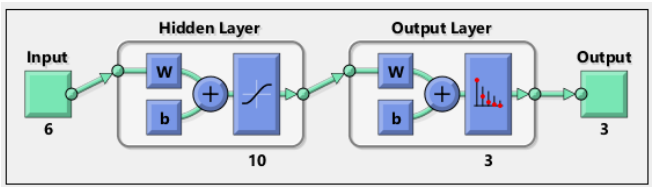

Figure 2. Neural network for Air Quality Index prediction

Results are shown in Fig. 3 and Table 2.

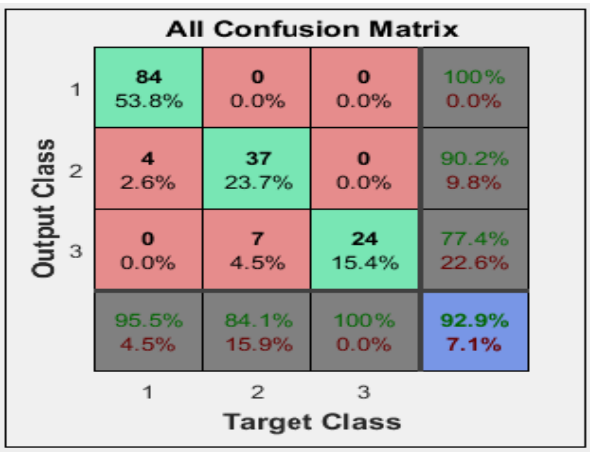

Figure 3. All Confusion Matrix when 70% Training, 10% Test and 20% Validation with highest accuracy of 92.9%

TABLE II. TABLE 2. ACCURACY OF NEURAL NETWORK WITH DIFFERENT VALUES FOR VALIDATION AND TESTING

Method	Data division	Accuracy
Neural network	70% train, 20% validation, 10% test	92.9%
Neural network	70% train, 15% validation, 15% test	90.4%
Neural network	70% train, 10% validation, 20% test	91.7%

IV. K- NEAREST NEIGHBOR ALGORITHM

k-NN is Supervised Learning classifier. k-NN are non-parametric techniques which are used for classification and regression [6, 7, 8]. Non-parametric means that it does not make any assumptions on the underlying data distribution. In both cases, input contains nearest neighbor test samples. Output depends on classification or regression:

In k-NN classification, output is member of the class. Object is classified according to the votes of most of the nearest neighbors, so that the object is assigned to a class which is most voted (k is positive integer, usually a small number). If k=1, then object is assigned to a class of its nearest neighbors.

In k-NN regression, output represents the value of the object. This value is an average of its values of nearest neighbors. k-NN is learning based on examples or lazy learning. Lazy learner means that it does not use training data to do any generalization. This means there is no explicit training phase or it is very small. Almost all training data are used during the testing phase.

Neighbors which are nearest to the object contribute more than neighbors that are far from the object. Neighbor has dedicated weight of 1/d, where d is a distance between neighbors. Neighbors are taken from the set of objects whose class is known. Weakness of the k-NN algorithm is that it is sensitive on local data structures. Most used distance metrics for continuous variables is Euclidean distance. There are many other metrics such as: city-block (sum of absolute differences); cosine (one minus the cosine of the included angle between points (vectors)); correlation (one minus the sample correlation between points (sequence of vectors)); hamming (percentage of bits that differ (suitable only for binary data)).

The best choice for k depends on data. In general, higher values of k are decreasing the effect of noise on classification, but they limit differences between classes.

A. *Experiments with k-NN Algorithm*

In k-NN algorithm few combinations for getting highest accuracy with different value of nearest neighbors (k) were done. Values are in interval of k=1 to k=21 and project contains 3 classification classes. Since there are only 2 classification classes, odd value for k is recommended [9, 10]. Testing was performed with different types of metrics: Euclidean, correlation, city block and cosine.

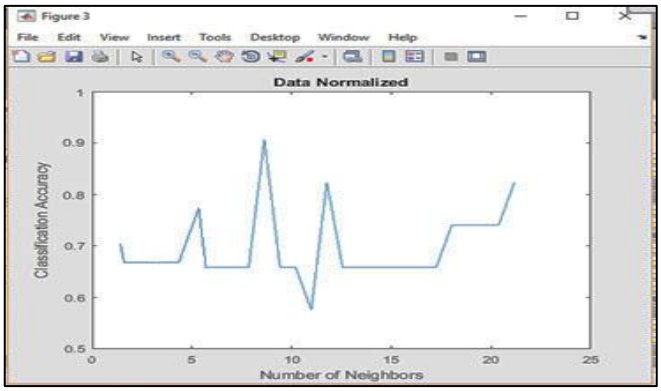

Figure 4. Values of k in interval {k=1, k=21}, when k=8, top accuracy is 90.5%

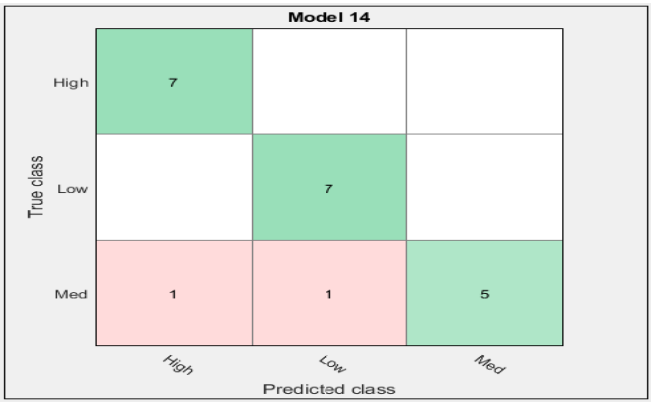

Figure 5. Confusion matrix when K=8

It can be seen from Figure 4, that the best accuracy for k-NN algorithm is for k=8. Figure 5 shows accuracy of algorithm in confusion matrix. Algorithm classifies high and low samples without error and in medium samples it classifies 5 true samples and 2 wrong, 1 in low and 1 in high air pollution class. Further tests will be performed for k=8 with different types of metrics.

Figure 6 and Table III show accuracy performed with different types of metrics.

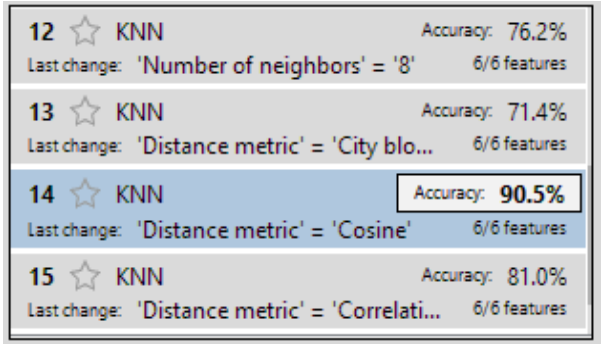

Figure 6. Accuracy of k-NN algorithm performing different types of metrics

TABLE III. ACCURACY OF K-NN WHEN USING DIFFERENT TYPE OF METRICS

Method	Processing time	Accuracy
k=8 | Euclidean | Equal	1.585 seconds	76.2%
k=8 | Correlation | Equal	1.542 seconds	81.0%
k=8 | City Block | Equal	1.524 seconds	71.4%
k=8 | Cosine | Equal	1.567 seconds	90.5%

From the table above, we can see that the algorithm has best accuracy when k=8 with cosine metrics.

V. DECISION TREE ALGORITHM

Decision tree is a supervised learning algorithm. It's one of the models which is used for statistic prediction, data mining and machine learning [11, 12]. Tree models where target variable contains set of discrete values, are called classification trees. In this tree structures, leaves are presenting class signs and branches are conjunctions of properties which lead to signs of the class. Decision trees where target variable get continuous values (typically real numbers) are called regression trees [13]. In decision analyses, decision tree can be used for visually or explicit presentation of decisions or used for making decisions. Purpose is to be created a model which will predict target value based on previously learnt input data.

Algorithms for constructing decision trees often work as steam (from the top to the end). Each internal node is test for some attribute, each branch presents result of the test (corresponds with value for tested feature) and each leaf is keeping the class mark (classification for an instance). Highest node of the tree is the root. Hierarchy of rules is implemented. Very rule of internal nodes (roots) tests value of some feature of the data. Training data are used for tree construction, then, depending on the tree model, output values are predicted. Information with highest value are located on the top of the tree.

Each internal node fits with one of the input variables. Each leaf is value of the target variables, according to input values, presented on path from the root to the leaves.

There are two types of decision tree algorithms [14]: classification tree (when predicted outcome is the class which contains the data and regression tree (when predicted result can be counted as real number). Classification and regression tree with one name are called CART (classification and regression tree).

Tree can be learnt with splitting the source data which are presented in subsets based on test characteristics values. This process is repeated on every of performed subgroups and it is called recursive partitioning. Recursion is done when subset in the node have the similar value of the target variable and when the splitting does not increase value of the predictions.

Many software packets for data mining are implementing decision trees. Some of them are: IBM SPSS Modeler, RapidMiner, SAS Enterprise Miner, Matlab, R, etc. [14].

A. *Experiments with Decision Tree Algorithm*

Decision tree is simple representation for sample classification [11, 13]. For this reason it is assumed that all input functions have final discrete domains and there is one target function called "classification". Each domain element of classification is called class. Each internal node in the classification tree is marked with input feature. Each leaf of the tree is marked with class.

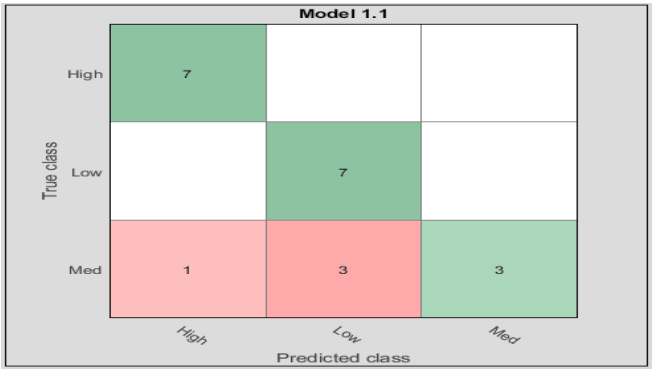

Figure 7. Confusion matrix for Decision tree algorithm

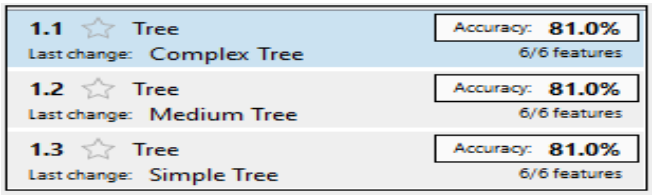

Figure 8. Accuracy of different sized Decision trees

Fig. 7 and Fig. 8 present the results of the accuracy of different sized decision trees: single tree, medium tree, complex tree. From the result on the figure above and the table 4, it can be seen that, for each decision tree size, accuracy result is same 81.0%. Just like in k-NN algorithm, DT experiments are made according to supervised learning, that means algorithm knows input and output values from training dataset and test dataset.

TABLE IV. ACCURACY OF DECISION TREES WITH DIFFERENT SIZES

Method	Number of splits	Accuracy
Single Tree	10	81.0%
Medium Tree	20	81.0%
Complex Tree	100	81.0%

VI. DISCUSSION OF THE RESULTS AND CONCLUSION

In this paper we have compared three different algorithms for data analysis: NN, k-NN and DT. The experiments were conducted with database of air pollution in Macedonia in 2016.

After the analysis of all the data, we came to a conclusion that the most accurate algorithm for analysis and classification of the result is NN with maximum accuracy of 92.9%, while k-NN algorithm has a maximum accuracy of 90.5% and DT algorithm has maximum accuracy of 81.0%.

NN algorithm contains 3 layers: input, 1 hidden and output layer. Input layer contains 6 input attributes, hidden layer contains 10 neurons and output layer contains 3 classes.

In the k-NN algorithm several combinations are made to obtain the highest accuracy with different values of the nearest neighbors (k). Values are in the interval when k=1 to k=21. This research leads to a conclusion that the greatest accuracy algorithm holds when k=8 and has a cosine metrics.

The decision trees are faster in data processing and easy to understand, but they are not as accurate as k-NN and NN.

It was noticed that classifiers DT and Back propagation NN look like "eager students", because they first build classification model based on training dataset, before they classify the test dataset.

Classifier based on k-NN, does not build classification model. It learns directly from training dataset. k-NN is processing data after test data is known to be classified, so, we can say it is "lazy student".

REFERENCES

[1] Christopher M. Bishop, "Neural Networks for Pattern Recognition", Oxford, 1995

[2] Nagendra, S.S.; Khare, M. "Artificial neural network approach for modelling nitrogen dioxide dispersion from vehicular exhaust emissions". Ecol. Model. 2006, 190, 99–115

[3] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, Oriol Vinyals, "Understanding deep learning requires rethinking generalization", ICLR 2017

[4] A., Nielsen, Michael (2015). "Neural Network and Deep learning", NeuralNetworksAndDeepLearning.com

[5] Bramer, M. "Principles of data mining", Springer-Verlag, London, 2013.

[6] Altman N.S. (1992). "An introduction to kernel and nearest-neihbor nonparametric regression"

[7] Samworth R.J. (2012). "Optimal weighted nearest neighbor classifiers".

[8] D. Coomans; D.L. Massart (1982). "Alternative k-Nearest neighbor rules in supervised pattern recognition: Part 1. k-Nearest neighbor classification by using alternative voting rules". Analytica Chimica Acta

[9] Veljanovska, K. "Machine Learning Algorithms Comparison", International Journal of Engineering and Technical Research (IJETR), Vol.7, No.11, 2017

[10] Veljanovska, K. "Analysis of Machine Learning Algorithms Performance for Real World Concept", International Journal of Advanced Research in Science and Engineering (IJARSE), Vol.06, No.12, 2017

[11] Mitchell, T. "Machine Learning", McGraw-Hill, New York, USA, 1997

[12] Sayali D. Jadhav, H.P. Channe, "Comparative Study of k-NN, Naïve Bayes and Decision Tree Classification Technologies", International Journal of Science and Research, 2013

[13] Kattariya Kujaroentavon, Supapom Kiattisin, Adisom Leelasantitham, Sotarat Thammaboosadee, "Air quality classification in Thailand based on decision tree", Biomedical Engineering International Conference (BMEiCON), 2014

[14] Breiman, Leo; Friedman, J.H; Olshen, R.A; Stone, C.J. (1984). "Classification and regression trees". Montrey, CA: Wadsworth & Brooks/Cole Advanced Books & Software

Prof. d-r Kostandina Veljanovska. D-r Kostandina Veljanovska **c**ompleted her education at the University "Sts. Kiril i Metodi", Skopje (BSc in Computer Science), at the University of Toronto, Toronto (MASc in Applied Engineering) and got her MSc and also her PhD in Technical Sciences at the University "St. Kliment Ohridski", Bitola, R. Macedonia. She has completed postdoc in Artificial Intelligence at the Laboratory of Informatics, Robotics and Microelectronics at the University of Montpellier, Montpellier, France. She worked as a Research assistant at the ITS Centre and Testbed at the Faculty of Applied Science, University of Toronto, Canada. She also, worked at research team for Constraints, Learning and Agents at LIRMM, University of Montpellier. Currently, she works as an Associate Professor in Information Systems and Networks, Artificial Intelligence and Systems and Data Processing at the Faculty of Information and Communication Technologies, University "St. Kliment Ohridski" -Bitola, Republic of Macedonia. Her research work is focused on artificial intelligence, machine learning techniques and intelligent systems. She has published numerous scientific papers in the area of interest, as well as several monographic items. She is a reviewing referee for well-known publishing house, journals with significant impact factor in science and also, member of editorial board of several international conferences.

Angel Dimoski. Angel Dimoski completed his education at the University "St. Kliment Ohridski", Bitola (BSc in Computer Science), in 2016. He currently works toward the MS degree in Information Sciences and Communication Engineering at University "St. Kliment Ohridski"- Bitola, Republic of Macedonia. His research interest includes artificial intelligence, robotics, virtual reality, engineer expert systems.