International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org
Volume 7, Issue 1, January - February 2018 ISSN 2278-6856

Air Quality Index Prediction Using Simple Machine Learning Algorithms

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

Abstract

Air pollution and its prevention are constant scientific challenges during last decades. However, they still remain huge global problems. Affecting human’s respiratory and cardiovascular system, they are cause for increased mortality and increased risk for diseases for the population. Many efforts from both local and state government are done in order to understand and predict air quality index aiming improved public health. This paper is one scientific contribution towards this challenge. We compare four simple machine learning algorithms, neural network, k-nearest neighbor, support vector machines and decision tree. The air pollution database contains data for each day of 2017 from measurements stations in capital city of the Republic of Macedonia. The results are promising and it was proven that implementation of these algorithms could be very efficient in predicting air quality index.

Keywords: Air Quality Index, Machine learning, Decision Tree, k-Nearest Neighbor, Neural Network, Support Vector Machine

1. INTRODUCTION

Artificial intelligence and machine learning are areas of biggest rise in the last year. The science of artificial intelligence where system makes decision on its own, instead of working only by orders given by programmer as traditional programming works, gradually started influencing all aspects of our life. Starting from early-stage startup companies and ending to large platform vendors, for all of them, artificial intelligence and its part machine learning have become the key focus areas.

Machine learning is an area where system which implements artificial intelligence gathers data from sensors in an environment and learns how to act. One of the reason why we choose machine learning to predict air quality index, was this ability of adapting of machine learning (ML) algorithms.

In this paper three supervised learning algorithms k-nearest neighbor (k-NN), Support Vector Machines (SVM) and Decision Tree (DT) and one unsupervised algorithm Neural Network (NN) are compared.

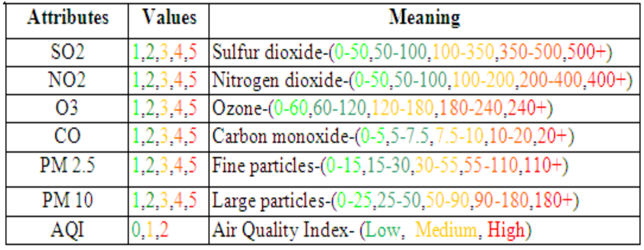
There are many researchers that implement some of the algorithms we are using, such as NN [1], NN and fuzzy systems [2], SVM [3], SVM for regression [4], fuzzy logic [5], DT [6], k-NN [7], but none of them compare their performance as one research for all of them four at the same conditions and for the same data.

2. AIR POLLUTION PROBLEM AND DATABASE CONSTRUCTION

Air pollution in the Republic of Macedonia has been concerned as a serious problem. Every day measured values of the parameters of air quality are many times above the limit values that are considered safe for human health. In the larger urban areas the situation is urgent. Some of the activities to lower the air pollution are undertaken by the local government, some by the state government. This paper represents our effort on the scientific level to contribute in handling this problem. At the first phase the necessary step is to predict air quality index in order to help improve the situation. In this project we developed four different classifiers based on different algorithms.

Database (Table 1) used in the project is based on model of official web site of Ministry of environment and physical planning of the Republic of Macedonia. Measurement taken by the stations include **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 1: Air pollution attributes (values are expressed in μg/m^3, and CO is expressed in mg/m^3)

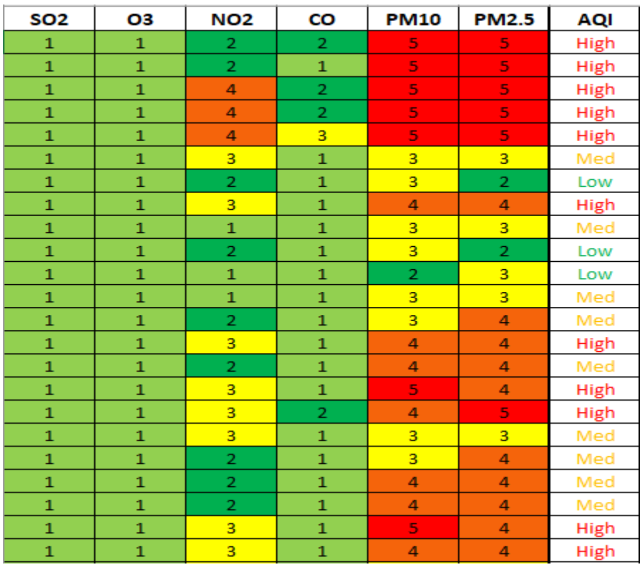


Attributes	Values	Meaning
SO2	1,2,3,4,5	Sulfur dioxide-(0-50,50-100,100-350,350-500,500+)
NO2	1,2,3,4,5	Nitrogen dioxide-(0-50,50-100,100-200,200-400,400+)
O3	1,2,3,4,5	Ozone-(0-60,60-120,120-180,180-240,240+)
CO	1,2,3,4,5	Carbon monoxide-(0-5,5-7.5,7.5-10,10-20,20+)
PM 2.5	1,2,3,4,5	Fine particles-(0-15,15-30,30-55,55-110,110+)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

As seen from Figure1, border of low/medium/high air pollution is set on index with value 3 (medium). Reviewing all six attributes, we set rules for the level of pollution. 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, independent on other attributes. If one attribute 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 database which is constructed according to air pollution in capital of the Republic of Macedonia, Skopje. The reason is because the major city of

Macedonia is the most polluted city in the world these days. Database contains 365 samples (each per day of 2017), 51 of this samples are with High Air Pollution Index, other 94 are with Medium Air Pollution Index and the rest 220 samples are with Low Air Pollution Index. Purpose of this project is to build four 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 Database contains 100 samples and it is used in k-NN, SVM and DT algorithms. Neural network works as unsupervised learning algorithm using same training database of 365 samples with 6 attributes. These samples are classified into 3 different classes: Low, Medium or High.



SO2	O3	NO2	CO	PM10	PM2.5	AQI
1	1	2	2	5	5	High
1	1	2	1	5	5	High
1	1	4	2	5	5	High
1	1	4	2	5	5	High
1	1	4	3	5	5	High
1	1	3	1	3	3	Med
1	1	2	1	3	2	Low
1	1	3	1	4	4	High
1	1	1	1	3	3	Med
1	1	2	1	3	2	Low
1	1	1	1	2	3	Low
1	1	1	1	3	3	Med
1	1	2	1	3	4	Med
1	1	3	1	4	4	High
1	1	2	1	4	4	Med
1	1	3	1	5	4	High
1	1	3	2	4	5	High
1	1	3	1	3	3	Med
1	1	2	1	3	4	Med
1	1	2	1	4	4	Med
1	1	2	1	4	4	Med
1	1	3	1	5	4	High
1	1	3	1	4	4	High

Figure 1 AQI according to attribute values

The experiments were performed in MATLAB Machine Learning Toolbox as a platform for experimenting.

3 Neural Network Modelling

Neural network is biological model which is interesting from data processing point of view since it calculates and adopts decisions and conclusions alike the human brain. One of the reasons why human is very good in solving processes like recognizing objects and faces, recognizing the environment, solving different situation problems, actions and reactions is the brain organization and functioning. From birth the brain has ability to develop its own rules based on earlier experience. These rules are made by strengthening old and making new connections between the neurons. Human brain has billions of neurons which are interconnected one another and communicate over electrochemical signals. Neural networks work like copy of the human brain to solve complex problems in machine learning area [8]. The learning in neural networks is not performed on each neuron particularly, but as a system, globally. All neurons learn together constructing a network which can solve problems with a high grade of accuracy. One of the most frequently used types of neural networks is Backpropagation NN where backpropagation is used to lower the error [9].

Backpropagation itself begins after calculation for input data ends. Activation function of each neuron is determined by the value of the output from the previous layer and weights between the neurons. First that need to be done is calculation of the error in the output result (distinction between the expected result and the actual result). Calculating the error is forward propagation. After the error is known with forward propagation, minimization is done using back propagation. This means that algorithm is propagating backwards, from output layer to input layer and on its way it finds error for each of the weights. That values have to be changed to minimize the total error [10].

3.1 Experiments with NN

Neural Network in this project is classifying samples to “low”, “medium” and “high” levels of air pollution, depending on the training database. 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 [11]. In our case, number of neurons in hidden layer is 10, because experimenting with this value, we get the smallest optimal error. In this project the neural network is constructed like unsupervised learning, which means that the input data for training are known to the network, but output data are unknown, so the network performs classification by knowledge gathered from input data.

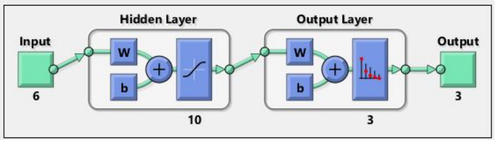
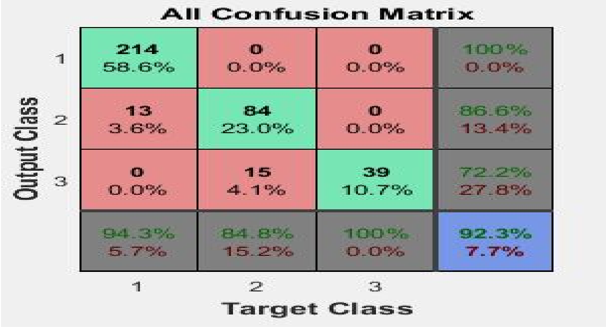


Figure 2 Neural Network Architecture for our model



All Confusion Matrix

Output Class \ Target Class	1	2	3	
1	214 58.6%	0 0.0%	0 0.0%	100% 0.0%
2	13 3.6%	84 23.0%	0 0.0%	86.6% 13.4%
3	0 0.0%	15 4.1%	39 10.7%	72.2% 27.8%
	94.3% 5.7%	84.8% 15.2%	100% 0.0%	92.3% 7.7%

Figure 3 All confusion matrix for the case with highest accuracy of 92.3%

4 K- Nearest Neighbor Algorithm Modelling

k-NN is Supervised Learning classifier. k-NN are non-parametric techniques which are used for classification and regression. If used for classification, result is object classification depending on the result of the nearest neighbors voting, and the new object is dedicated to the class with more votes. On the other side, if used for regression, then the result is value that is dedicated to the object as an average of the values of its neighbors [12, 13, 14].

In k-NN algorithm the training phase is not performed. Unlike the other algorithms, k-NN algorithm does not make any presumptions for data distributions, and, also, does not bring general conclusions. Because there is not training phase, it must keep all the data for training and search over them for neighbors. This process requires more time and more resources for testing phase.

There are many distance metrics in k-NN. The most used distance metrics is Euclidean distance if continuous variables are considered. Other metrics that are often used are: city-block, hamming distance, cosine and correlation.

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

4.1 Experiments with k-NN algorithm

The experiments were done with few combinations in order to get the highest accuracy taking into account different value of nearest neighbors (k). Values of k were in interval from 1 to 21. Project has 3 classes for classification. Odd value for k was taken [15, 16]. Testing was performed with several types of metrics: Euclidean, correlation, city block and cosine.

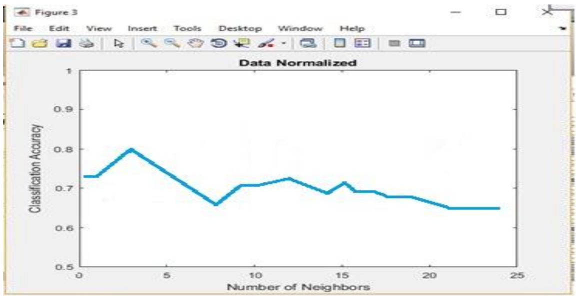
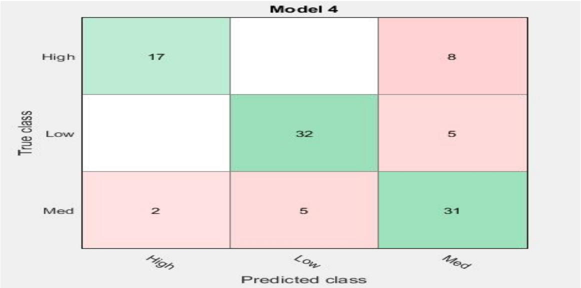


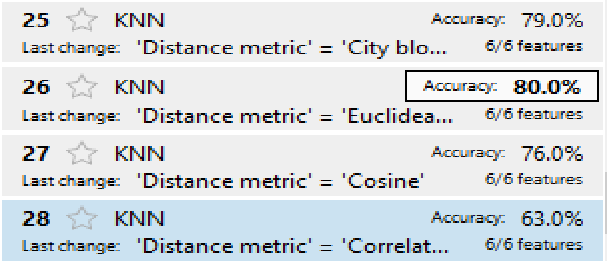
Figure 4 Values of k and classification accuracy



	High	Low	Med
High	17		8
Low		32	5
Med	2	5	31

Figure 5 Confusion matrix when k=3

Analyzing the experiments, it can be seen that the best accuracy is for k=3 (Figure 4). Figure 5 shows confusion matrix of accuracy of k-NN algorithm. Algorithm classifies samples as "high level of pollution" as follows, 17 correct and 8 in wrong (in medium class), "low level of pollution": 32 correct and 5 wrong (in medium class) and in class of "medium level of pollution" 31 correct samples while others belong to high and low class. Further tests will be performed for k=3 with different types of metrics.



25 KNN Accuracy: 79.0%
Last change: 'Distance metric' = 'City blo... 6/6 features

26 KNN Accuracy: 80.0%
Last change: 'Distance metric' = 'Euclidea... 6/6 features

27 KNN Accuracy: 76.0%
Last change: 'Distance metric' = 'Cosine' 6/6 features

28 KNN Accuracy: 63.0%
Last change: 'Distance metric' = 'Correlat... 6/6 features

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

k-NN algorithm has best accuracy when k=3 with Euclidean metrics (Table 3).

5 DECISION TREE MODELLING

Decision tree as a supervised learning algorithm is used as a model for statistic prediction, data mining and machine learning [17, 18]. There are two types of decision tree algorithms [19]: classification tree (predicted outcome is the class which contains the data) and regression tree (predicted result is real number). Classification and regression tree with one name are called CART (classification and regression tree).

Classification trees are tree models where target variable contains set of discrete values. In this structures, leaves are presenting class signs and branches are conjunctions of properties which lead to signs of the class. Regression trees are decision trees where target variable gets continuous values (typically real numbers) [20]. Decision trees are used to make decisions and to visually or explicitly present them. The aim is to create a model which will predict target value based on previously learnt input data.

Decision tree algorithm is constructed as a steam where each internal node represents test for an attribute, each branch represents value for tested feature and each leaf is classification for an instance. Highest node of the tree is the root. Hierarchy of rules has to be implemented. Rules of internal nodes (roots) test values of some feature of the data. 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.

Training data has to be used for tree construction. Afterwards, depending on the tree model, output values are predicted. Information with highest value are located on the top of the tree.

The process of learning the tree can be performed by splitting the source data presented in subsets based on test characteristics values. This splitting should be repeated on every of the subgroups and it is called recursive partitioning. At the moment when subset in the node have the similar value of the target variable and when the splitting does not increase value of the predictions recursion is finished.

Some of the software packets for data mining implementing decision trees are: IBM SPSS Modeler, RapidMiner, SAS Enterprise Miner, Matlab, R, etc. [19].

5.1 Experiments with Decision Tree algorithm

In order to perform experiments, DT was constructed with assumption that all input functions have final discrete domains, there is one target function classification of air pollution data (three classes). Every internal node in the DT contains input feature. Every leaf of the tree contains class.

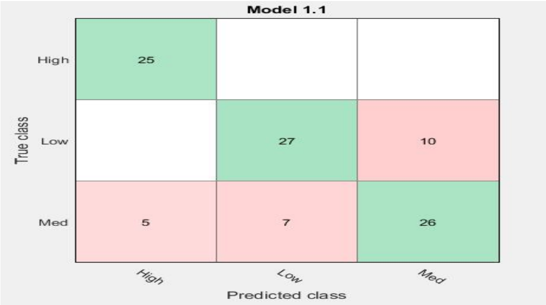
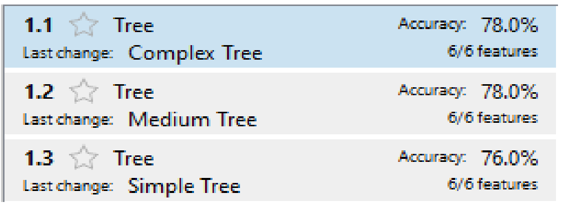


Figure 7 Confusion matrix for DT algorithm


1.1 ☆ Tree Accuracy: 78.0%
Last change: Complex Tree 6/6 features

1.2 ☆ Tree Accuracy: 78.0%
Last change: Medium Tree 6/6 features

1.3 ☆ Tree Accuracy: 76.0%
Last change: Simple Tree 6/6 features

Figure 8 Accuracy of different sized DT

The result of the accuracy of different sized decision trees: single tree, medium tree, complex tree are presented on Figure 7, Figure 8 and Table 5. It can be seen that, for complex and medium tree, the accuracy result is same 78.0%, and for simple tree is 76.0%. Just like in k-NN algorithm, DT and SVM, experiments were made as supervised learning, i.e. algorithm knows input and output values from training database and test database.

6 Support Vector Machines Modelling

Kernel methods are a class of machine learning techniques that has become an increasingly popular tool for learning tasks such as pattern recognition, classification or novelty detection [21]. This popularity is mainly as a result of the success of the support vector machines (SVM), probably the most popular kernel method, and to the fact that kernel machines can be used in many applications as they provide a bridge from linearity to non-linearity.

The fourth model that we choose for comparison in our project is support vector machines**.** Support vector machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Some of the advantages of SVM are [22]: effective in high dimensional spaces; different kernel functions can be used for various decision function; kernel functions can be combined to achieve more complex type of planes, although SVM have poor performance when number of features is grater then number of samples and SVM do not provide probability estimates which is the reason why cross-fold validation is used.

SVM model is a representation of the examples as points in space, mapped in such a way that the examples of the separate categories are divided by a major vector (hyper plane) which is as wide as possible. Left and right from that major vector, supports vectors at the same distances from major vector are positioned. New examples are then mapped into the same space and predicted to belong to a category based on that on which side of the vector they fall. So, the result depends on the position of the major vector. This is called linear support vector machine (LSVM) [23].
Except performing linear classification, SVMs can, also, efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. Using kernel function, two vectors can be applied and every point is mapped into high dimensional space by a transformation. The idea is to transform non-linear space into linear space [24].
There are several popular kernel types that can be used to transform data into high dimensional feature space: polynomial kernel, radial basic function kernel, sigmoid kernel.

6.1 Experiments with SVM algorithm

Figure 9 shows that 20/25 High samples are classified as correct and 5 of them are classified in Medium class. 31/37 Low class samples are correctly classified and 6 of them are classified in Medium class. And 29/38 samples of Medium class are correctly classified, 9/38 classified as wrong, 4 of them are in Low class and 5 samples in High class.

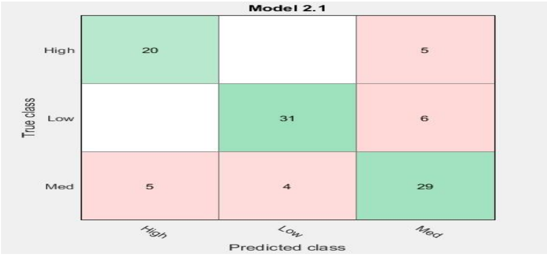
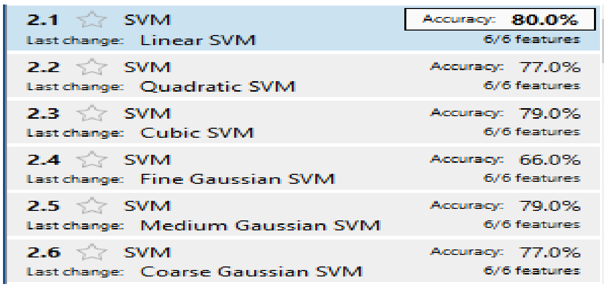


Figure 9 Confusion matrix of highest accuracy of SVM algorithm



2.1 ☆ SVM Accuracy: 80.0%
Last change: Linear SVM 6/6 features

2.2 ☆ SVM Accuracy: 77.0%
Last change: Quadratic SVM 6/6 features

2.3 ☆ SVM Accuracy: 79.0%
Last change: Cubic SVM 6/6 features

2.4 ☆ SVM Accuracy: 66.0%
Last change: Fine Gaussian SVM 6/6 features

2.5 ☆ SVM Accuracy: 79.0%
Last change: Medium Gaussian SVM 6/6 features

2.6 ☆ SVM Accuracy: 77.0%
Last change: Coarse Gaussian SVM 6/6 features

Figure 10 Accuracy of SVM algorithm using different kernel functions

From the figures 9 and 10 and Table 4, it can be seen that SVM algorithm have highest accuracy of 80.0%, when using Linear kernel function.

7 Discussion of the Results and Conclusion

The paper compares four different algorithms for machine learning: NN, KNN, DT and SVM. The experiments were conducted using database of air pollution in the capital, Skopje, of the Republic of Macedonia in 2017.

NN algorithm contains one input, one hidden and one output layer. Input layer contains 6 input attributes, hidden layer contains 10 neurons and output layer contains 3 classes. Table 2 shows the accuracy of NN for the experiments performed for different number of data for training, validation and testing. It was found that the most accurate case is NN when 70% of data are used for training, 10 % for validation and 20% for training the NN with maximum accuracy of 92.3%.

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	91.8%
Neural network	70% train, 15% validation, 15% test	91.5%
Neural network	70% train, 10% validation, 20% test	92.3%

For the k-NN algorithm several combinations were made to obtain the highest accuracy with different values of the nearest neighbors (k). Values were taken in the interval from k=1 to k=21. Table 3 shows the accuracy of the algorithm for different type of metrics. This research leads to a conclusion that the greatest accuracy algorithm holds when k=3 and has Euclidean metrics.

Table 3: Accuracy of k-NN for different type of metrics

Method	Processing time	Accuracy
k=3 | Euclidean | Equal	1.711 seconds	80.0%
k=3 | Correlation | Equal	1.794 seconds	63.0%
k=3 | City Block | Equal	1.851 seconds	79.0%
k=3 | Cosine | Equal	2.197 seconds	76.0%

In SVM algorithm different kernel functions were tried to get highest accuracy result. Experiments lead to a conclusion that maximum accuracy of SVM was when linear kernel function was used. Table 4 shows the results for SVM with different kernel functions.

Table 4: Accuracy of SVM with different kernel function

Kernel Function	Accuracy
Linear	80.0%
Quadratic	77.0%
Cubic	79.0%
Fine Gaussian	66.0%
Medium Gaussian	79.0%
Coarse Gaussian	77.0%

The decision trees are faster in data processing and easy to understand, but they are not as accurate as others K-NN, SVM and NN. Table 5 presents the accuracy for different DT sizes.

Table 5: Accuracy of Decision trees with different sizes

Method	Number of splits	Accuracy
Single Tree	10	76.0%
Medium Tree	20	78.0%
Complex Tree	100	78.0%

After the analysis of the results, it can be concluded that the most accurate algorithm for classification is NN with maximum accuracy of 92.3%, while KNN algorithm has a maximum accuracy of 80.0%, DT algorithm has maximum accuracy of 78.0% and SVM algorithm has maximum accuracy of 80.0%.

References

[1] Q. Feng, “Improving Neural Network Prediction Accuracy for PM$_{10}$ Individual Air Quality Index Pollution Levels”, Environmental Engineering Science, 30(12): 725–732, 2013

[2] P. Hájek, V. Olej, “Prediction of Air Quality Indices by Neural Networks and Fuzzy Inference Systems – The Case of Pardubice Microregion”, International Conference on Engineering Applications of Neural Networks (EANN), pp 302-312, 2013

[3] L. Wang, Y.P. Bai, “Research on Prediction of Air Quality Index Based on NARX and SVM”, Applied Mechanics and Materials (Volumes 602-605), 3580-3584, 2014

[4] BC. Liu, et al, “Urban air quality forecasting based on multi-dimensional collaborative Support Vector Regression (SVR): A case study of Beijing-Tianjin-Shijiazhuang”, PLOS, 2017

[5] H. Wang, et al., “Air Quality Index Forecast Based on Fuzzy Time Series Models”, Journal of Residuals Science & Technology, Vol. 13, No. 5, 2016

[6] N. Loya, et al., “Forecast of Air Quality Based on Ozone by Decision Trees and Neural Networks”,

Mexican International Conference on Artificial Intelligence (MICAI), pp 97-106, 2012

[7] E. G. Dragomir, “Air Quality Index Prediction using K-Nearest Neighbor Technique”, BULETINUL Universităţii Petrol – Gaze din Ploieşti, Vol. LXII No. 1/2010 103 – 108, 2010

[8] C. M. Bishop, “Neural Networks for Pattern Recognition”, Oxford, 1995

[9] C. Zhang, et al., “Understanding deep learning requires rethinking generalization”, ICLR, 2017

[10] M.A. Nielsen, “Neural Network and Deep learning”, Determination Press, 2015

[11] M. Bramer, “Principles of data mining”, Springer, 2007

[12] N.S. Altman, “An introduction to kernel and nearest-neihbor nonparametric regression”, The American Statistician, Vol. 46, No. 3, pp. 175-185, 1992

[13] R.J. Samworth, “Optimal weighted nearest neighbor classifiers”, 2012

[14] D. Coomans; D.L. Massart, “Alternative k-Nearest neighbor rules in supervised pattern recognition: Part 1. k-Nearest neighbor classification by using alternative voting rules”. Analytica Chimica Acta, 1982

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

[16] K. Veljanovska, “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

[17] Mitchell, T. “Machine Learning”, McGraw-Hill, New York, USA, 1997

[18] J.D. Sayali, H.P. Channe, “Comparative Study of k-NN, Naïve Bayes and Decision Tree Classification Technologies”, International Journal of Science and Research, 2013

[19] L. Breiman, et al., “Classification and regression trees”. Montrey, CA: Wadsworth & Brooks/Cole Advanced Books & Software, 1984

[20] K. Kujaroentavon, et al., “Air quality classification in Thailand based on decision tree”, Biomedical Engineering International Conference (BMEiCON), 2014

[21] S. Canu, “SVM and kernel machines: linear and non-linear classification”, OBIDAM, Brest, 2014

[22] A. Suárez Sánchez et al., “Application of an SVM-based regression model to the air quality study at local scale in the Avilés urban area (Spain)”, Mathematical and Computer Modelling 54, pp. 1453–1466, 2011

[23] C. Cortes, V. Vapnik, “Support-vector network”, Machine Learning, 1995.

[24] B. Gaonkar, C. Davatzikos, “Analytic estimation of statistical significance maps for support vector machine based multi-variate image analysis and classification”, 2014.

AUTHORS

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.