

# Comparison of Different Machine Learning Algorithms for Air Quality Predictions

International University of Sarajevo, Faculty of Engineering and Natural Sciences, HrasnickaCesta 15, Ilidža71210 Sarajevo, Bosnia and Herzegovina

Student: Međuseljac Almir

Professor: Emine Yaman

**Abstract**

Weather is very big factor in everyday life. In addition to the fact that weather affects our mood, it also affects our health. There are several factors that affect what the weather will be like, for example geographical location, infrastructure, population, etc. Sarajevo is a capital city from Bosnia and Herzegovina and like most other capital cities in the region, in winter time it faces problems with smog level in the air. This problem is allready „common“ thing that is happening from long time ago. Unfortunately,every year from november to mid-january, the citizents of Sarajevo face the problem of air pollution every day. It is not strange for the inhabitants of this beautiful city if they town ranks first on the world list of the most polluted cities. Goverment is still trying to find the solution to this problem, to reduce index of pollution and PM10 particles in the air.

Further in the paper I will appy several machine-learning algorithms like KNN algorithm and J48 tree. KNN algorithm, as the words say k-nearest neighbors, is a simple lazy machine learning algorithm that can solve classification and regression problems. J48 (C4.5) ispopular algorithm which is used to generate decision tree, and the developer of this algorithm is Ross Quinlan. C4.5 generates decision trees that can be used for classification, this is reason than C4.5 is often referred to as a static classifier.

1. **Introduction**

Weather within a state or a citiy is affected by many factors, so it is very important. Man is able to affect weather to some extent, thus in what relationship he has with he nature. Of course, when everything is taken into consideration, there is lot of factors affecting the weather. In this paper I have singled out one weather disaster that is common occurrence in Sarajevo.Also I have collected certain information on air pollution in Sarajevo and did analysis of that same data. The data set that I have created contains several attributes. First attribute is date. In this data set I collected informations from 27. November 2020. untill 27. December 2020..Further in the data set I collected informations for each day individualy. Those measurements collected were taken every day at 10am. Second attribute is index of pollution. The index of pollution depends on PM2.5 particles that have a diameter of less than 2.5 micrometers, which makes them 100 times thinner than human hair. These particles are result of the combustion of various fuels and chemical reactions that take place inside the atmosphere. These small particles, which are mainly the product of various combustion, due to theis size very easily manage to reach lungs and then the bloodstream to reach many internal organs, so they have great impact on human health. The amount od PM2.5 particles can be devided into several sections: from 0 to 12.0 – good, 12.1 to 35.4 – modernate, 35.5 to 55.4 – unhealthy for sensitive groups, 55.5 to 150.4 – unhealthy, 150.5 to 250.4 – very unhealthy, 250.5 to 500.4 hazardous. Pm10 particles are a very common occurence when it comes to air pollution. Their diameter is 10 micrometers or even smaller. It can be found in dust or smoke. Air density is also attribute of our dataset, and it is calculated mass per unit volume of Earth's atmosphere. Last attribute in data set is healthy, which indicates whether given conditions are healthy for humans or not.

# 

1. **Literature review**

There are many researches on the amount of smog in Sarajevo, but in this paper i will compare results between the two machine-learning algorithms. First algorithm will be J48 decision tree, second algorithm will be KNN algorithm.

# 

1. **Material and Methods**

This data set contains apecific weather informations. In the data set we have 3 numeric attributes and 2 binary or nominal-valued attributes. We have attributes such as date, index of pollution, index of Pm10 particles, air density and healthy. First method that I am going to implement here on my data set is J48 decision tree. Decision trees are one of the basic procedures of machine learning and it is a popular procedure because of simplicity of execution and especially because the result of indication can be graphically displayed. In figure 1. we can see J48 decision tree obtained using Weka software for 31 day smog prediction.

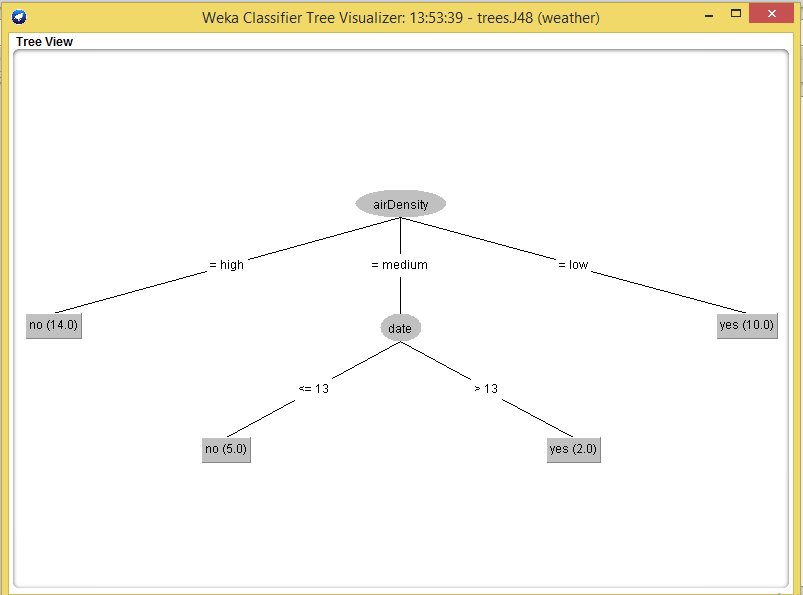


Figure 1.

In figure 1. at the root node we have the air density from which the branches „=high“, the branch „=low“ and the branch „=medium“ emerge. In the case of „=low“ the branch immediately ends with an end node because all examples that have „=low“ density are in the same class. The class of this end node is „yes“ and the data that follows is (10.0). Also with the branch „=high“ we reach the end node. The class of this node is „no“ and the data that follows is (14.0). With the „=medium“ branch, we come to the decision node where the „date“ attribute is. From this attribute we have two branches. First one i „<=13“ which ends in the final node „no“ with the characteristic (5.0). In this node we can find those days in which „airDensity = medium“ & „date <=13“ and the number of them is 5. The second branch is „>13“ and it ends in the final node „yes“ with the characteristics (2.0). This means that those nodes that have „airDensity =medium“ & „date >13“ end in this node and there are 2 of them.

After using the 10-fold cross-validation, we came to some rules:

*if airDensity = high: no*

*if airDensity = medium and if date <=13 : no*

*if airDensity = medium and if date >13 : yes*

*if airDensity = low : yes*

These are our rules that we got when applying the J48 decision tree machine learning algorithm. In his algorithm, we tried to classyfy 31 instances, 28 of which were identified correctly and 3 that were identified incorrectly. This gives us a percentage of 90.3226% correctly identified and 9.6774% identified incorrectly. The average precision in our data set is 0.904 and our average recall is 0,903. In the figure2. I will show you how to calculate the precision and recall.

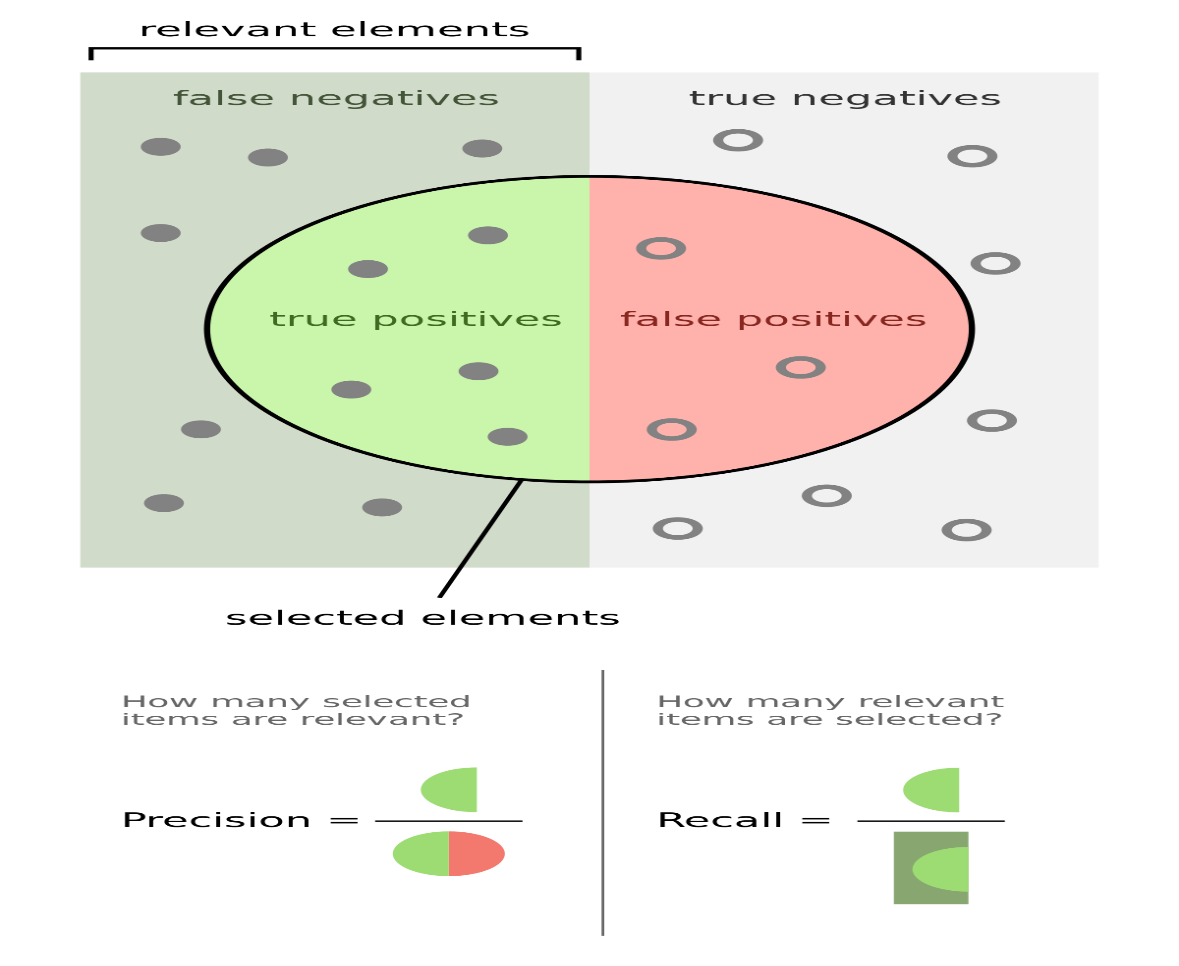


Figure 2.

Second algorithm that I have used for data mining is KNN machine learning algorithm. KNN or K nearest neighbour in Weka is a lazy method, that can be used to solce both classification and regression problems.Letter k in KNN stands for number of nearest neighbours to include in the voting proces, since KNN stores and classifies new cases based on similarity measure. Since we will be talking about Euclidean Distance, in figure 3 I will provide formula for the calculation of Euclidean Distance.

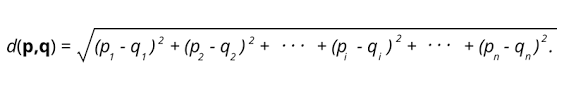


Figure 3.

In my dataset I have used IBk algorithm two times. In first algorithm i was calculating using the Euclidean Distance with K=1. Results that i got showed that out of 31 iinstances, 29 of them were classified correctly, 93.5884% and 2 instances were incorrectly classified, 6.4516%.

Average weighted precision for this algorithm was 0.935 and average recall was 0.935. In the second step i increased the number of K to K=2 and i got slightly different results. Again, by using Euclidean Distance, our correctly identifyed instances were same, 29 out of 31 which makes it to 93.5484%. Incorectly classified instances were 6.4516% which also stayed the same.We notice a small diference in precision which was 0.942 in this case and our recall was 0.934 for K=2.

This machine learning algorithm is very accurate but slow. Also this algorithm is assuming that all attributes are equally important, and this k-NN algorithm have been in use ince 1950s.

**Findings**

The tables below contain data on the results of the various machine learning algorithms we applied to the data set.

|  |  |  |
| --- | --- | --- |
| Algorithm /  Attribute-healthy | Correctly Classified Instances | Incorrectly Classified Instances |
| **IBK** | 29  93.5484% | 2  6.4516% |
| **RandomForest** | 28 90.3226% | 3 9.6774% |
| **J48** | 28 90.3226% | 3 9.6774% |

|  |  |  |
| --- | --- | --- |
| Algorithm / Attribute airDensity | Correctly Classified Instances | Incorrectly Classified Instances |
| **IBK** | 28 83.871% | 5 16.129% |
| **RandomForest** | 28 90.3226% | 3 6.6774% |
| **J48** | 26 83.871% | 5 16.129% |

**Conclusion**

Air quality control is very important within a city like Sarajevo. Especially in winter where air pollution is higher than in the rest of the year. I did this research with the aim of gaining more knowledge about this topic. Also, this research served educational purposes, with the aim of getting to know better data mining tools and techniques and to see their applications in real life.

**References:**

*(1) Principles of Data Mining   
 by M.A. Bramer and Max Brame*

*(2) Data Mining Practical Machine  
 Leatning Tools and Techniques  
 by Ian H. Witten, Eibe Frank  
 Mark A. Hall, Christopher J. Pal*

*(3) Machine Learning Algorithm for   
 Data Science and Machine Learning   
 Giuseppe Bonaccorso, 2017*

*(4) https://zrak.ekoakcija.org/sarajevo  
 ?order=epaaqi&sort=asc*

*(5) https://ba.usembassy.gov/embassy/  
 sarajevo/air-quality-monitor-sarajevo/*