Data Mining Algorithms for Weather Forecast Phenomena：Comparative Study

Marwa Farouk M.Ali, Somia A. Asklany, M. Abd El-wahab, M.A.Hassan

Egyptian Meteorological Authority, Cairo, Egypt.
Northern Border University, Faculty of science and arts, Girls section, Computers department, Turaif, K.S.A
Cairo University, Faculty of science, Department of Astronomy and Meteorology, Giza, Egypt.
Ain Shames University, Faculty of science, Mathematics department, Cairo, Egypt.

Summary
In Meteorological field, where a huge database takes place; weather prediction is a vital process as it affects people's daily life. In the last century, the accuracy of weather predictions has been one of the most challenging concern facing meteorologists around the world. Atmospheric dust is considered to be a harmful air pollutant causing respiratory diseases and infections from one side as well as affecting the earth's energy budget from the other side, so an early prediction of dust phenomena occurrence can be very useful in reducing its harmful effects. Data mining is mainly a machine learning process for extracting useful information form extremely large data base as it is capable of handling huge, noisy, ambiguous, random and missing data, so it represents a very helpful tool in predicting different weather elements. The virtue of using data mining techniques is that they not only analyse the huge historical data base, but also learn from it for future predictions. In this work, we investigate the use of data mining techniques in forecasting different atmospheric phenomena specially atmospheric dust using Decision Tree, k-NN and Naïve biased algorithms as well as making a comparison between them by evaluating each model results. The proposed models are implemented using the open source data mining tool Rapidminer.

Key words:
Data Mining, Decision Trees, Naïve Biased, KNN, rapid miner.

1. Introduction

Meteorology as a science concerned with the study of the atmosphere, weather phenomena's and climate issues. The observation of change in the weather elements such as temperature, air pressure, moisture and wind direction usually owns a large sets of databases as any of those weather elements is measured and recorded every hour. Databases are rich with hidden information that can be used by forecasters to understand certain metrological phenomena and make important decisions that may help in saving peoples life. Forecasters need some new methods to extract this useful information from the large databases.

The major reason that data mining has attracted a great deal of attention in information industry in recent years is due to the wide availability of huge amounts of data and the very demanding need for turning such data into useful information and knowledge [1]. Data mining is a very useful technique not only for data classification but also for prediction process.

Data mining proved its ability to be a very helpful tool for weather state prediction such as rainfall [2,3] , thunderstorm, maximum temperature, evaporation, wind speed [4] and cloud burst [5]. Other researchers even used data mining for guiding the path of the ships during sailing [6].

Atmospheric dust is composed of a mixture of solid and liquid particles suspended in the atmosphere varying in composition, source and size. Suspended dust particles affect radiative balance of the earth system [7].

Those dust particles can be removed out (washed out) of the atmosphere by a process called deposition and fall back on soil, vegetation or watercourses. The deposition process means that the aerosols particles are collecting or depositing in a solid surfaces. Dust particles in the atmosphere are deposited onto the Earth's surface by solution in precipitation (wet deposition) or without precipitation (dry deposition) [8]. Dust particles can be classified according to their diameter, measured in micrometres, as

Primary particles – diameter from 2.5 to 30 µm;

Secondary particles – diameter lower than 2.5 µm.

Primary particles are form by the natural sources of dust while Secondary particles are generated by human activities such as vehicular traffic and industrial activities. PM10 and PM2.5 is defined to be atmospheric dust particles with a diameter of less than 10 µm and 2.5 µm respectively, those particles draw a special attention for their serious impact on human health [9]. For example if the percentage of PM10 is above its normal levels determined by World Health Organization (WHO) for human health protection with is 20 µg/m3 as an annual average and 50 µg/m3 as a daily average, dust particles can get into the lungs casing wheezing, chest tightness and difficulty breathing, irritation of eyes and throat. People with existing heart or lung disease (including asthma) can experience a serious increase in symptoms [10].Wind speed also plays an important role in either increasing or decreasing the air pollutant concentrations (PM10

Manuscript received September 5, 2019
Manuscript revised September 20, 2019

concentration). For instance, the low wind speed produces weak dispersion of the pollutants which causes weak ventilation and the pollutants are found at the highest concentrations. However, strong wind speeds in places of arid or semi-arid lands like Sahara desert in Africa can form dust storms by blowing up the dust particles on the ground [11]. Dust particles can also affect the strength of visibility [12] due to the various particle compositions and their interactions with light, low visibility could be an indicator of high PM concentrations.

According to World metrological organization (WMO), the four main categories dust phenomena that can reduce visibility are:

Haze: (which is suspension of dry particles of dust in the atmosphere), reduces the horizontal visibility to less than 5KM, with a very calm wind (the wind speed is less than 2.5m/s).

Blowing Dust: reduces the horizontal visibility to less than 5KM.

Dust Storm: reduces the horizontal visibility to less than 1 Km.

Severe Dust Storm: the horizontal visibility is less than 500 m.

Several Researches [13– 15] have estimated the contribution of Sahara desert alone by 2 x $10^8 - 3.3$ x 10^8 tons every year or between 40–66% of the total dispersed dust. When tracing Dust storms, it was found that it may travel to a distance about 4000 km from their origin. For Egypt dust storms are frequent weather phenomena occur in all seasons of the year. Last year only Egypt has been attacked by Dust storms four times, Driven mainly by west or south west wind, they result from the erosion and transport of mineral sediments from the ground surface of the desert part of east Libya.

In this work we used some important metrological variables such as wind speed, visibility, relative humidity and the PM_{10} concentration to predict different types of atmospheric dust using data mining techniques decision tree, KNN and Naive bias algorithms under rapid miner software platform.

1.1 Decision Tree

Decision Tree represents a decision support tool very often used because it is simple to understand and interpret. [16]

It is flow-chart-like tree structure. Each internal node denotes a test on an attribute. Each branch represents an outcome of the test. Leaf nodes represent class distribution. After generation, the decision tree model can be applied to new Examples to make predictions, so in order to classify an unknown sample, the attribute values of the sample are tested against the decision tree. A path is traced from the root to a leaf node which holds the class prediction for that sample, each Example follows the branches of the tree in accordance to the splitting rule until a leaf is reached. [1]

Different algorithms can be used for building decision trees such as CHAID (Chi-squared Automatic Interaction Detection), CART (Classification And Regression Trees), Quest and ID3 (Iterative Dichotomiser 3), which is the precursor to the C4.5 algorithm, and is typically used in the machine learning and natural language processing domains. [17]

So to create a decision tree ,the following steps should be done

1. Create root node for the tree.
2. If all examples are positive, return leaf node ‘positive’.
3. Else if all examples are negative, return leaf node ‘negative’.
4. Calculate the entropy of current state H(S).
5. For each attribute, calculate the entropy with respect to the attribute ‘a’ denoted by H(S,a).
6. Select the attribute which has maximum value of information gain (IG).
7. Remove the attribute that offers highest IG from the set of attributes.
8. Repeat the above steps until we test all attributes, or the decision tree has all leaf node.

Suppose that S be a set consisting of s data samples, Suppose also that the class label attribute has n distinct values defining n distinct classes, C_i (for i = 1,...…, n). Let s_i be the number of samples of S in class C_i (class label). The expected information needed to classify a given sample is given by:

$$H\left(s_{1,}s_2, \dots\dots\dots, s_n\right) = -\sum_{i=1}^{n} P_i log_2(P_i) \qquad (1)$$

Where P_i is the probability than an arbitrary sample belongs to class C_i, $log_2(P_i)$ is a function used since the information is encoded in bits.

H is called the entropy and the entropy is a measure of disorder or impurity of every set.

The entropy can take one of those three values,

If Entropy = 1; it means that the positive attributes = negative attributes. {impure set}

If Entropy = 0; it means that the set contains only positive attributes or negative attributes. {Pure set}

If Entropy is a value between 0 and 1; it means that the positive attributes and negative attributes contained in some set is not equal.

Note that for only two elements $\{s_1, s_2\}$ in the class C_i, then the entropy is defined to be

$$H(S) = -P_+ \log(P_+) - P_- \log(P_-) \qquad (2)$$

where P_+ is the probability of positive attribute and P_- is the probability of negative attribute.

By calculating the Information gain to get the attribute that has the highest information gain to be chosen as the test attribute for the given data set S. A node is created and labelled with the attribute, branches are created for each value of the attribute, and the samples are partitioned accordingly.

$$\text{IG(S, A)} = \text{H(S)} - \text{H(A)} \qquad (3)$$

$$\text{IG(S, A)} = \text{H(S)} - \sum_{i=1}^{v} P(a_i)\, log_2\, P(a_i) \qquad (4)$$

Where A is a Let attribute A have v distinct values, A=$\{a_1, \dots \dots \dots, a_v\}$. Attribute A can be used to partition S into v subsets, $\{S_1, \dots \dots \dots, S_v\}$.

1.2 KNN Algorithm

KNN algorithm is considered to be one of the most popular machine learning algorithms, it has been used in statistical estimation and pattern recognition already in the beginning of 1970's as a non-parametric technique, it can be used for both classification and regression predictive problems.

In this method the missing values of an attribute are computed using a given number of attributes that are most similar to the attribute whose values are missing. The similarity of two attributes is determined using a distance function.

Different metrics, such as Euclidean distance can be used to calculate the distance between the unknown Example and the training Examples.

For calculating the Euclidean distance between two points, X = $\{ x_1, x_2, \dots \dots, x_n\}$ and Y= $\{ y_1, y_2, \dots \dots, y_n\}$,

$$\text{d(X,Y)} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad (5)$$

equation(5) refers to the fact that distances often depends on absolute values, so if a set contains some variables of different measurement scales or has a mixture of numerical and categorical variables, like one variable is based on annual income in dollars, and the other is based on age in years then income will have a much higher influence on the distance calculated, and the results is not accurate. So for better results it is recommended to normalize data, i.e. to scale all values for a given attribute so they fall within a small specified range, such as -1.0 to 1.0, or 0 to 1.0 before training and applying the k-Nearest Neighbor algorithm [1].

The normalization equation

$$x_n = \frac{x - x_{min}}{x_{max} - x_{min}} \qquad (6)$$

Where x_n is the normalized value of x.

KNN algorithm has the advantage that it can predict both numerical and categorical attributes and Attributes with multiple missing values can be easily treated, also the Correlation of the data is taken into consideration.

However when analyzing large database, the algorithm is very time-consuming, as It searches through all the dataset looking for the most similar instances. Choosing the value of k is very critical for example, higher value of k may include attributes very different from the required one and it will be computational expensive. Whereas lower value of k implies missing out of significant attributes, i.e. the noise will be higher.

1.3 Naive Bayes

Bayesian classifiers are statistical classifiers. They can predict class membership probabilities, such as the probability that a given sample belongs to a particular class. Bayesian classification is based on Bayes theorem which states that

$$P(X|H) = \frac{P(H|X)}{P(X)} \qquad (7)$$

Where X is a data sample whose class label is unknown, and H is some hypothesis, such as that the data sample X belongs to a specified class C. For classification problems, we want to determine P$(H|X)$, the probability that the hypothesis H holds given the observed data sample X.

The performance of the naive Bayesian classifiers can be compared with decision tree and neural network classifiers. Bayesian classifiers also have the advantage of high accuracy and speed when applied to large databases.

Naive Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called class conditional independence [1].

1.4 Rapidminer

Rapidminer is an open and extensible data science software platform developed by the company of the same name that provides an integrated environment for data preparation, machine learning, deep learning, text mining, and predictive analytics. It has so many applications in different fields such as business, commercial applications, scientific research, education, training, rapid prototyping, and application development and supports all steps of the machine learning process including data preparation, results visualization, model validation and optimization. Rapid miner also provides an advanced analytics at every scale perfect for big data, Strong Visualizations, Multiple Interfaces and Accurate Pre-processing. [18]

2. Methodology

2.1 Data Collection

The data used for this work was collected from Cairo Airport reports, those reports contain the recorded hourly data of all-weather elements which are pressure, Temperature, Relative humidity, dew point, wind speed, wind direction and outlook. The data collected is covering a period of five years, from January 2010 to December 2015. And the data of the year 2017 is used as a test data. The method flow chart is shown in figure 1.

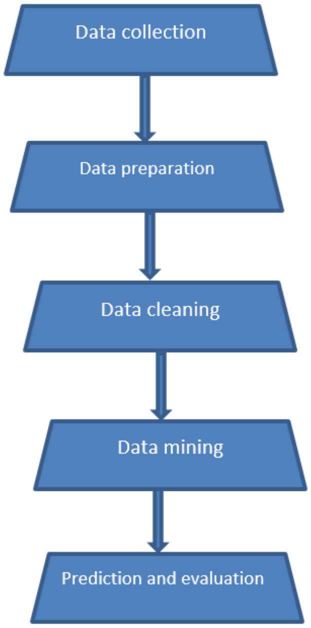


Fig. 1 proposed method flow chart

2.2 Data Preparation

Selected data sets were transformed into a format of excel sheet to be suitable for the mining process.

2.3 Data Cleaning

In this stage, after, the missing data, duplicated data, and outliers data are filtered and excluded from Data set. The cleaned data were obtained and prepared for the data mining stage.

2.4 Data Mining

At the data mining stage we divide our data set into two sets: One set with no missing values for the variables and another one with missing values. First data set is considered to be the training data set on which we apply the algorithm to build the model. The other data set with missing values is test data set and variables with missing values are treated as target variables to test the model performance. The testing method used in this research was percentage split that train the model on a certain percentage of the dataset, cross validate on it and test on the remaining percentage. At the end of this process interesting patterns representing knowledge were extracted and identified.

3. Experiments and Results

In this work three experiments were done, through each one we performed analysis, simulation and forecasting performance evaluation for the leaning algorithms using the Rapid miner version 9.0 software.

In the first experience we build a model with decision tree algorithm for predicting the different weather state (outlook), in the second one we also build a model with decision tree algorithm for predicting different categories of dust phenomena, and the third experience was made to compare performance of the three algorithms, decision tree, KNN and naïve biased .

The confusion matrix, correlation and root mean square are used to evaluate different models.

3.1 The First Experiment

In table 1, we gave a full description for the attributes used in the experiment.

Table 1: Attribute Description

Attribute	Description
Wind speed	Measured by m/s and it is classified as follows From 0 : 4 calm wind From 5: 7 moderate From 8 : 10 Active Above 10 : Storm.
Visibility	Measured by meters
Relative humidity	Percentage
Outlook	Haze (HZ) , Dust (Du) ,dust storm (DS) ,Cavok (clear weather case),fog (FG) and mist (BR)

In Fig. 2, the decision tree diagram made by rapidminer after entering our required data.

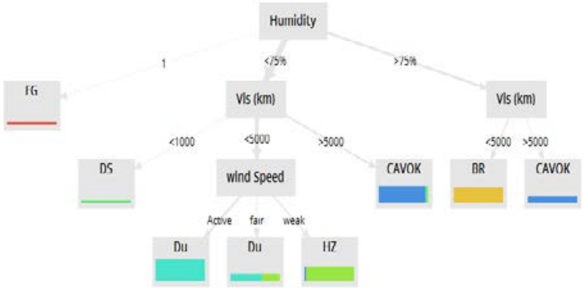
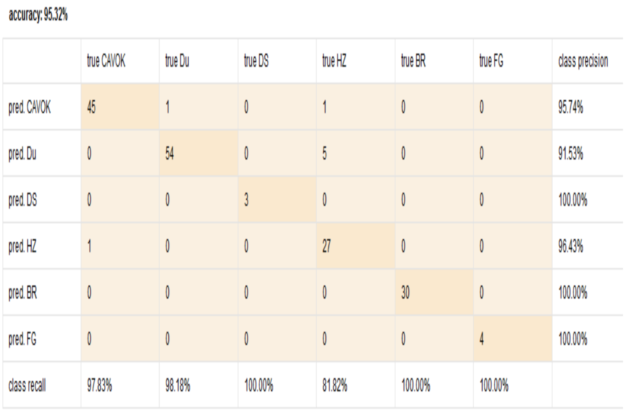


Fig. 2 decision tree diagram

In fig. 2, the confusion matrix, which is an essential step for model evaluation, is a table form used to describe the performance of a classification model on a set of test data, its diagonal represent the numbers of true predictions while the other off diagonal numbers represent the false predictions, the accuracy of the true predictions of the model is calculated by the equation

$$Acurracy = \frac{the\ sum\ of\ the\ numbers\ on\ the\ diagonal}{sum\ of\ the\ whole\ numbers\ in\ the\ matrix} \quad (8)$$

also table 2 gives the results of Model accuracy and correlation as another step for the model evaluation.



accuracy: 95.32%

	true CAVOK	true Du	true DS	true HZ	true BR	true FG	class precision
pred. CAVOK	45	1	0	1	0	0	95.74%
pred. Du	0	54	0	5	0	0	91.53%
pred. DS	0	0	3	0	0	0	100.00%
pred. HZ	1	0	0	27	0	0	96.43%
pred. BR	0	0	0	0	30	0	100.00%
pred. FG	0	0	0	0	0	4	100.00%
class recall	97.83%	98.18%	100.00%	81.82%	100.00%	100.00%	

Fig. 3 confusion matrix

Table 2: model evaluation

Model Accuracy	95.32%
Root mean square error	0.190
correlation	0.954

3.2 The Second Experiment

We built a model using decision tree algorithm to classify different dust phenomena as well as predict each of them, using three attributes, see table 3. The figure 3 represented the full model building process illustration.

Table 3: attributes description

Attribute	Description
Wind speed	Measuered by m/s
Visibility	Measured by meters
Outlook	Haze (HZ) , Dust (Du) ,Sand storm (DS) , Cavok (for the clear weather case).

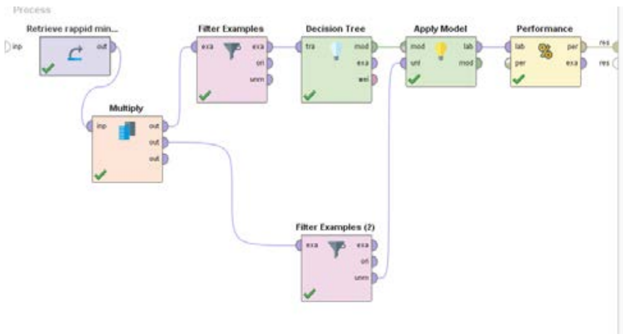


Fig. 4 displaying model in rapidminer

The decision tree for the case was

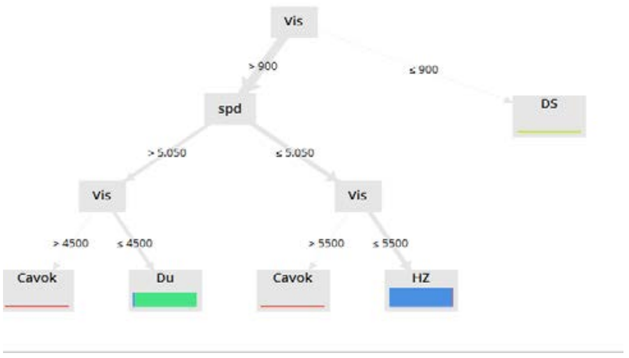
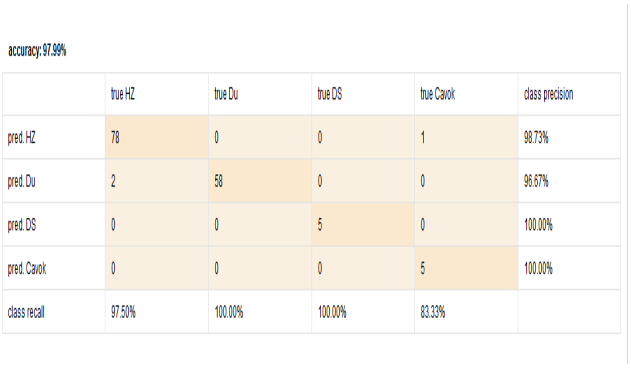


Fig. 5 the decision tree diagram

The model validation step by calculating the confusion matrix, Model accuracy, correlation and root mean square error. (See table 4).



accuracy: 97.99%

	true HZ	true Du	true DS	true Cavok	class precision
pred. HZ	78	0	0	1	98.73%
pred. Du	2	58	0	0	96.67%
pred. DS	0	0	5	0	100.00%
pred. Cavok	0	0	0	5	100.00%
class recall	97.50%	100.00%	100.00%	83.33%	

Fig. 6 the confusion matrix

Table 4: model evaluation

Model Accuracy	97.99%
Root mean square error	0.166
correlation	0.931

3.3 The Third Experiment

Predicting atmospheric dust phenomena by using pm10 concentration and building three models using decision tree algorithm, naïve biased and KNN algorithms, then comparing results.

In table 5 attributes desecration for the same range of data mentioned above.

Table 5: attributes description

Attribute	Description
Wind speed	Measured by m/s and it is classified as follows From 0 : 4 calm wind From 5: 7 moderate From 8 : 10 Active Above 10 : Storm
Pm10 concentration	Measured by $\mu g/m^3$
Outlook	Haze (HZ) , Sand (Du) ,Sand storm (DS) .

Figure 7, represents the Decision tree model one of the models used in the experiment.

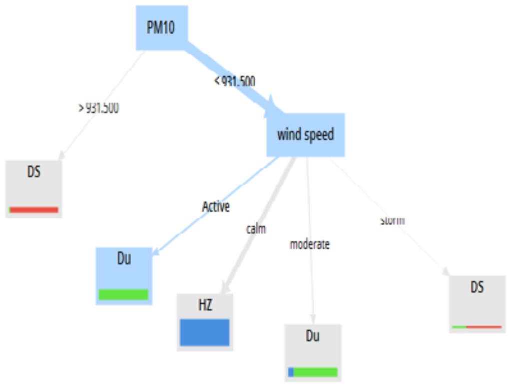


Fig.7 decision tree diagram

The performance of the three models was evaluated, and the results were illusterated in table 6.

Table 6: compression between three models

	Decision tree	KNN	Naïve biased
Accuracy	97.45%	77.34%	97.45%
Root mean square error	0.148	0.365	0.143
correlation	0.972	0.744	0.972

4. Conclusion

In this work we have performed experiments and compared data mining algorithms including Naive Bayes, KNN and Decision tree for weather forecast phenomena. The results portrayed that Decision tree more successful in classifying and modelling Data set it also proved its effectiveness in both classification and perdition. The behaviour of KNN algorithm was the weakest among the three algorithms. Naive Bayes which is a simple classifier based on Bayes theorem, is a simple classifier to apply and proves to be efficient in performance against the other two classifiers used as it gives nearly the same results of the Decision tree algorithm.

References

[1] Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers,San Francisco (2005)

[2] Casas D. M, Gonzalez A.T, Rodrígue J. E. A., Pet J. V, ,"Using Data-Mining for Short Term Rainfall Forecasting", Notes in Computer Science, 2009, Volume 5518, 487-490 .

[3] Somia A. Asklany, Khaled Elhelow , I.K. Youssef, M. Abd El-wahab, "Rainfall events prediction using rule-based fuzzy inference system, Atmospheric Research 101, 2011 228–236.

[4] Kaya, E.; Barutçu, B.; Menteş, S. "A method based on the van der Hoven spectrum for performance evaluation in prediction of wind speed". Turk. J. Earth Science, 2013, 22, 1–9.

[5] K. Pabreja, R.K. Datta, "A data warehousing and data mining approach for analysis and forecast of cloudburst events using OLAP-based data hypercube", Accepted by International J. Data Analysis Techniques and Strategies, Inderscience Publishers , 2012 ,Vol.4 , 57-82 ,

[6] P.Hemalatha, "Implementation of Data Mining Techniques for Weather Report Guidance for Ships Using Global Positioning System", International Journal Of Computational Engineering Research, march 2013, Vol. 3 Issue. 3.

[7] Dentener, F.J.; Carmichael, G.R.; Zhang, Y.; Lelieveld, J.; Crutzen, P.J. Role of mineral aerosol as a reactive surface in the global troposphere. J. Geophys. Res. 1996,101, 22869–22889. [CrossRef].

[8] Lovett, G. M.: Atmospheric deposition of nutrients and pollutants in North America: An ecological perspective, Ecol.1994, Appl., 4, 629– 650.

[9] Schwartz, J.; Dockery, D.W. Increased mortality in Philadelphia associated with daily air pollution concentrations. Am. Rev. Respir. Dis. 1992, 145, 600–604.

[10] Lin M, Chen Y, Burnett RT, Villeneuve PJ, Krewski D. Effect of short-term exposure to gaseous pollution on asthma hospitalisation in children: a bi-directional case-crossover analysis. J Epidemiol Commun H. 2003;57(1):50–55. doi: 10.1136/jech.57.1.50. [PMC free article] [PubMed] [CrossRef]

[11] Natsagdorj, L.; Jugder, D.; Chung, Y.S. Analysis of dust storms observed in Mongolia during 1937–1999.

[12] Appel, B.R.; Tokiwa, Y.; Hsu, J.; Kothny, E.L.; Hahn, E. Visibility as related to atmospheric aerosol constituents. Atmos. Environ. (1967) 1985, 19, 1525–1534.

[13] Junge C: The importance of mineral dust as an atmospheric constituent. in Morales C (ed):Saharan Dust, Scope 14, New York, John Wiley & Sons, 1979, pp. 49–60.

[14] Morales C: Saharan Dust. Scope 14, New York, John Wiley & Sons, 1979, pp. 297.

[15] Ganor E, Mamane Y: Transport of Saharan dust across the eastern Mediterranean. Atmos Environ 1982;16:581–587.

[16] Lior Rokach, Oded Maomom,"Data Mining with Decision Tree: Theory and Application", World scientific publishing Co. Pte Ltd., 2008.

[17] Herbert A. Edelstein, Introduction to Data Mining and Knowledge Discovery, Third Edition. (1999)

[18] www.rapidminer.com .