Applying Data Mining to Measured Values of Spring Water Parameters to Determine the Relationship between Them

[1]M. Fatih Adak, [2]Cem Sen, [3]Ibrahim Cil

[1]Department of Computer Engineering, Faculty of Computer and Informatics, Sakarya University, Turkey
[2]Department of Electrical and Electr. Eng., Fac. of Engineering, Sakarya University, Turkey
[3]Department of Industrial Engineering, Faculty of Engineering, Sakarya University, Turkey

*Corresponding author: M. Fatih ADAK Address: Department of Computer Engineering, Faculty of Computer and Informatics, Sakarya University, Turkey. E-mail address: fatihadak@sakarya.edu.tr, Phone: +902642957049

Abstract

The quality measurements of drinking water sources are crucial, and the values measured should be checked carefully. The water values are measured at specific intervals at drinking water treatment plant in Kocaeli, Turkey, and a treatment process is performed if the values are not at desired levels. In this study, data mining is applied to data obtained from these measurements, and using the results obtained, it is intended to find a relationship among the attributes. According to results, Chlorine has the biggest effect on Turbidity. With a decision support system to be developed, one can check the values and decide whether the water will be treated or not.

Key words: drinking water; gini algorithm; treatment plant; water quality

1. Introduction

Water, the source of life, is of profound importance in our life. Since life cannot be considered without water, the conservation and control of drinking water sources are vital. In this study, data mining is applied to the data obtained from the measurements carried out at drinking water treatment plant in province of Kocaeli, Turkey, and it is intended to form a relationship among these values.

The Directorate General of ISU (Kocaeli, Turkey Water and Sewer System Administration) is an institution which distributes water to water distribution network and serves about 550,000 subscribers via more than 500 pumping stations and water tanks in a total of 11 towns within the province of Kocaeli in its area of responsibility. It meets the water need of the province of Kocaeli, with a population of some 1,601,720 people. Due to the health issues, the institution must work quite sensitively in the distribution of water and the measurement of quality parameters of the water distributed. For this purpose, the SCADA system was established in order to distribute water, operate it, carry out real-time measurements and check the area instruments from the center. Within the scope of this system, the quality parameters (pH, chlorine, and turbidity) of water arriving in water tanks from the sources are measured in real time, and instant changes arrive in central computers of SCADA, where retrospective records can be kept as well. These measurement values can be kept for 5 years retrospectively, and although a quite large databank is formed, the retrospective data saved in the system cannot be used. Of these data, only instant changes are taken into consideration in the institution, and the

Doi: 10.5505/apjes.2013.76486

water supplied to the network is intervened if necessary only according to instant changes.

When the measured for water quality parameters of turbidity and pH have changed, either the valves are turned off or the amount of chlorine applied to water is adjusted. The values of these water quality parameters are provided in Table 1.

Table 1. Limit values for the measured parameters of water quality

Name of Parameter	Limit Values	Quality
Turbidity	≤ 1 NTU	Good
	1 NTU ≤ and ≤ 5 NTU	Moderate
	≥ 5 NTU	Bad
pH	6.5 ≤ and ≤ 9.5	Standard
	7.5	Ideal
Chlorine	0.2 ≤ and ≤ 0.5	Ideal

1.1. pH

pH is a logarithmic measurement that denotes the acidity or basicity of water. It represents the concentration of H+ ions found in the solution.

Distilled water is in balance in terms of H+ and OH- ions, and its pH value is 7. pH can be measured depending on the electric potential of H+ ions or with color indicators (e.g. phenolphthalein) [1].

1.2. Cholorine

It shows the concentration of chloride ions in water. Chloride is a type of ion found considerably widely in all natural or used waters [1]. Chlorine is an important parameter in water quality and the amount of chlorine should be set well [2, 3, 4].

1.3. Turbidity

Turbidity is a measurement of light permeability of water containing suspended solids. The reason for turbidity may be anything from suspended substances to large visible sediment in water [5]. Substances such as sand, clay, silica, calcium carbonate, iron, manganese, and sulfide cause turbidity.

Turbidity, which is high in river waters, particularly results from the soil brought by rain or from the domestic-industrial wastewater mixing with the river [6]. In addition, during this contamination, inorganic substances mix with water as much as organic substances do. The presence of these substances supports the formation of bacteria in water [7].

Today the computer and computer-aided information systems have become part of everyday life. A work used to last for hours before is possible to end in seconds today, thanks to this technology. Likewise, computer-aided information systems facilitate settlement of problems encountered in management decisions and shorten the time required. The involvement of a country in developing world society depends on its becoming more efficient, its use of time in the most efficient way, and its taking of prospective measures depending on the changing conditions. It can do so by acting more rationalistically and quickly when making strategic decisions by evaluating all alternatives thoroughly based on different scenarios.

Managers spend the majority of their time taking decisions. Taking accurate and consistent decisions requires the production of necessary information, which is possible through devising systems that supply the information. Immediately after the real-time measurement of various parameters on a location basis thanks to the sensors placed at the stations on the drinking water distribution network within the SCADA system operated within the Directorate General of ISU, these values arrived in the central computers of SCADA with time tags and the retrospective data were saved.

Many studies have been made regarding determination of water quality by focusing on assessment parameters. Astel et al. applied the non-hierarchical K-means classification algorithm on data set of chemical indicators of river water quality [8]. In a study by Zhang et al., data mining was applied to data set in which the measured values of quality parameters were collected, and a decision tree that decides whether the water was of good or bad quality was built.

The parameters they considered were conductivity, the amount of dissolved oxygen, pH, temperature, and turbidity [9]. In another previous study, Lu et al. built a decision tree by using data mining methods to predict raw water quality. They tested this decision tree with the values the simulator produced. With the decision tree they built, data were entered and the results were checked. It was seen that the results complied well with the actual results [10]. Another real time classification study on water quality has been done by [11]. In their study, Karimipour et al. (2005) carried out quality management of water by using data mining and the geographic information system (GIS). They investigated the question of what quality water was found in what region by means of the GPS and the values measured.

The parameters they considered included pH, DO (the amount of dissolved oxygen), and BOD (the biological oxygen demand). They examined the amounts of increase and decrease in each value according to population density in the region to assess on the quality of water [12].

The work carried out in this study is mainly related the examination of data saved in the databank of SCADA and the formation of retrospective characteristic information. To decide on quality of water coming from the source depending on the characteristic of time-dependent amount of water and the rainy and dry weather conditions affecting water's turbidity. A Decision Support System (DSS) was developed in which the parameter values measured with the SCADA system were used. The devised DSS A data set is created by taking records on different days and in different months from the data recorded by the water treatment plant. So, it is thought that it can reflect all conditions. The data set contains 517 records. An example of the data set created can be is a system based entirely on measured and real values of the parameters through the comparison of current and retrospective values of some parameter measurements performed simultaneously on the drinking water distribution network.

In the proposed system, past and current data were utilized to make fair comments. In this system, it is aimed to provide people with high quality water through controlling essential quality indicators stemming from different weather conditions. One of the benefits of this devised system is to ensure that the necessary measurements are taken in advance and that decision scenarios regarding the new investments or the working methods are constructed.

Since the measurements are performed at very frequent intervals, even a daily record consists of great data. Hence, during forming the data set, data were collected for different months and days. The various conditions of the seasons and rain were also taken into consideration in order for the data to reflect all conditions.

Data mining, classification, and the Gini algorithm, an algorithm to build a decision tree, were applied to the data set collected, and 3 different decision trees were obtained. These decision trees will result in putting forward and idea about, with the two attributes being known.

2. Data and Method

seen in Table 2. The data set consist of measured values of chlorine, pH and turbidity and also the dates.

Table 2. An example from the data set

Chlorine	pH	Turbidity	Date
0.21	7.7	0.52	01.11.2012 00:11
0.21	7.71	0.53	01.11.2012 00:25
0.21	7.71	0.53	01.11.2012 00:39
0.21	7.7	0.53	01.11.2012 00:53
0.2	7.7	0.52	01.11.2012 01:07
0.2	7.7	0.53	01.11.2012 01:21
0.2	7.7	0.51	01.11.2012 01:35
0.2	7.71	0.51	01.11.2012 01:49
0.2	7.71	0.51	01.11.2012 02:03

The Gini algorithm was used for classify the data set. By making use of this algorithm, 3 different decision trees are built in which the result sets are turbidity, pH and chlorine, respectively. The process of the study is in Fig 1.

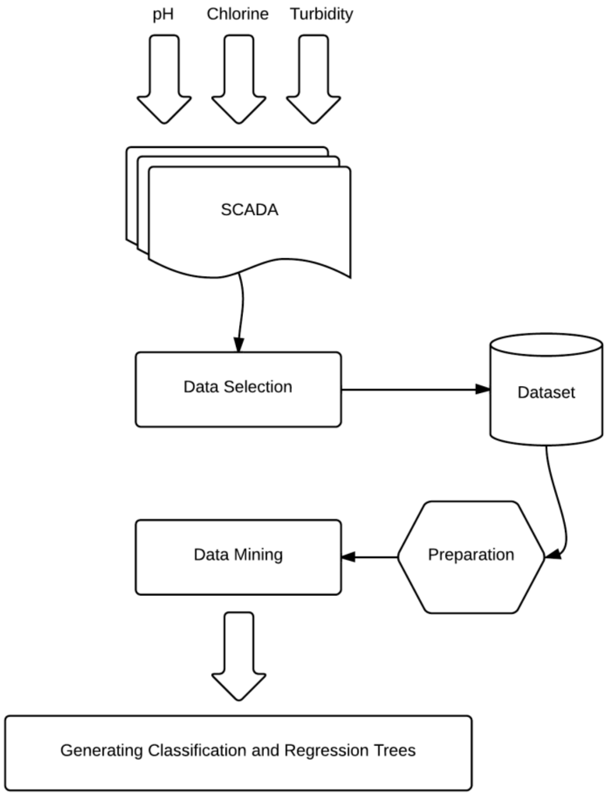

Figure 1. The process of generating trees

2.1. Data mining and classification

In order to perform data mining, data should be available at hand. Data mining mostly depends on statistical computations. It has various methods such as classification, clustering, and decision tree building.

Data mining can be defined as finding unexpected relationships among an observational data set, generally of a large volume by analyzing it, and making the data set understandable and useful by using novel methods [13].

Classification and decision-making, two important methods of data analysis, may determine essential data classes or build constructive models. In this way, these methods can enable predictions and comments on the future data. Classification is used to predict categorical data. Techniques such as decision trees, artificial neural networks, and genetic algorithms are generally used in classification [14, 15, 16].

2.2. Gini Algorithm

The Gini algorithm is an algorithm which is used to build a decision tree. It is defined as the CART. The CART (Classification and Regression Trees) decision tree is based on the principle of dividing the decision tree into two branches for each decision node [17]. Attributes will be divided first and the division value is computed considering the Gini index value. Gini index value can be defined as the rate of entities in the data set. If Gini values of two entities turn out to be the same, it means that their result distributions are the same. If there are 3 or more options in an attribute in the data set and since not more than two divisions are allowed, those options which are close to each other are grouped. An example is provided in Fig 2.

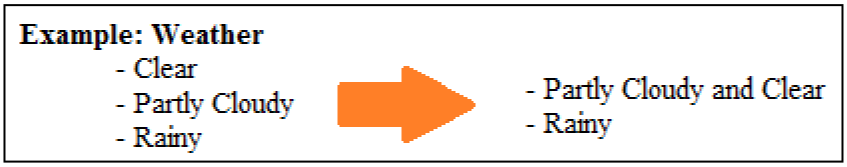

Figure 2. An example of division of the attribute

The Gini algorithm does not give successful results in all type of data sets and in some cases; it may be fail to construct the tree. In this case, one or several of the following stopping rules can be applied.

- If the node has become pure,
- If all nodes have become pure,
- If the tree has reached the maximum depth,
- If the minimum node size has been reached,

Before computing the Gini value of an attribute, the Gini left and Gini right values of the attribute should be computed. These computations are performed as in Equations 1 and 2. Meanings of the symbols in equations are given in Table 3.

$$\mathrm{Gini}_{\mathrm{left}} = 1 - \sum_{i=1}^{k} \left[\frac{L_i}{|T_{\mathrm{left}}|} \right]^2 \qquad \text{Eq. (1)}$$

$$\mathrm{Gini}_{\mathrm{right}} = 1 - \sum_{i=1}^{k} \left[\frac{R_i}{|T_{\mathrm{right}}|} \right]^2 \qquad \text{Eq. (2)}$$

Table 3. Symbols in the Gini equations

Symbol	Meaning
k	Number of classes
T	Examples in a node
T_{left}	Number of examples on the left-hand side
T_{right}	Number of examples on the right-hand side
L_i	Number of examples in category i on the left-hand side
R_i	Number of examples in category i on the right-hand side

The left and right values computed are used to compute the Gini value of the attribute. The computation of the Gini value is shown in Equation 3.

$$\mathrm{Gini}_j = \frac{1}{n}(|T_{\mathrm{left}}|\mathrm{Gini}_{\mathrm{left}} + |T_{\mathrm{right}}|\mathrm{Gini}_{\mathrm{right}}) \qquad \text{Eq. (3)}$$

The smallest one among the Gini values computed for each attribute is selected, and division takes place on this attribute. The above-mentioned steps are reapplied to the remaining data set, and the other division is computed.

3. Results

The Gini algorithm is applied to the data set by using the SPSS Clementine program, and 3 different decision trees are obtained. The decision tree giving an idea about what value the turbidity parameter might have when the parameters of chlorine and pH are known can be seen in Fig 3.

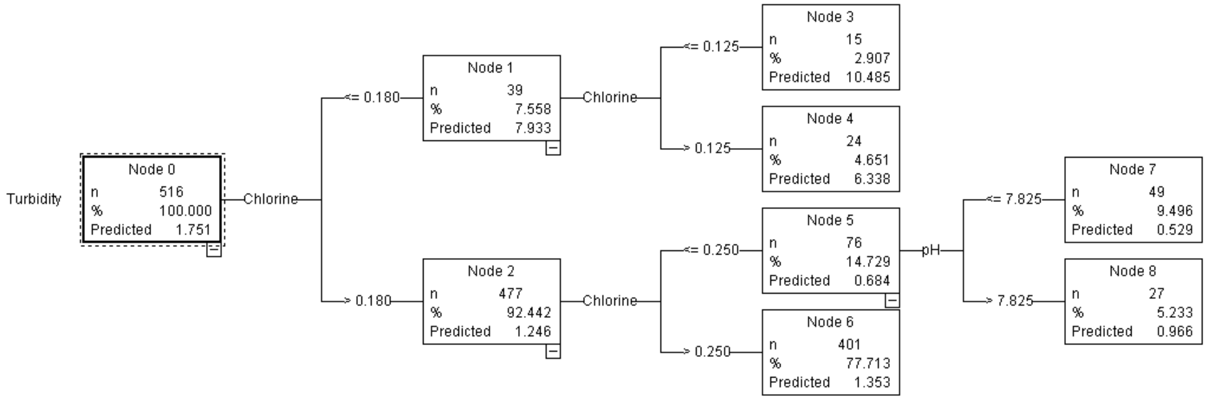

Figure 3. The decision tree in which the result set is turbidity

As can be seen from Fig 3, turbidity is high for the cases in which chlorine is greater than 0.18, whereas pH is inadequate to take a decision about turbidity.

The effect of chlorine on the decision tree is seen in Fig 4.

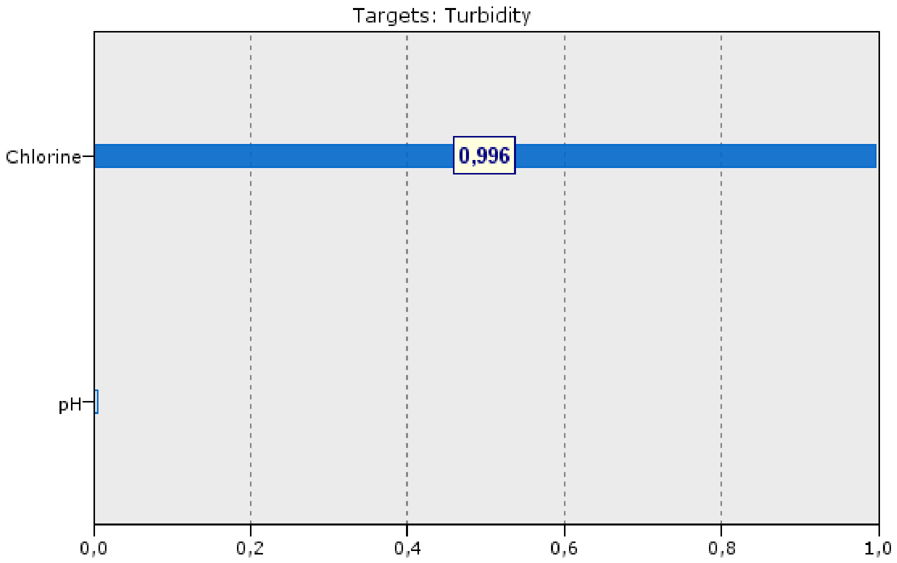

Figure 4. The effect of chlorine and pH on turbidity

The decision tree giving an idea about what value chlorine might have when the parameters of turbidity and pH are known can be seen in Fig 5.

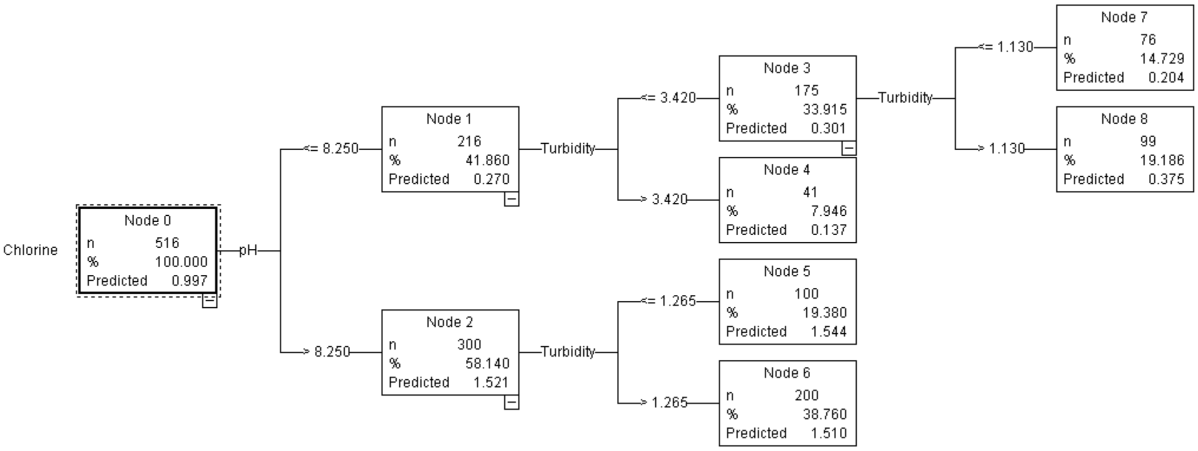

Figure 5. The decision tree in which the result set is chlorine

As can be seen from Fig 5, chlorine diverges from ideal values when pH is greater than 8.25. The decision tree giving an idea about what value pH might have when the parameters of turbidity and chlorine are known can be seen in Fig 6.

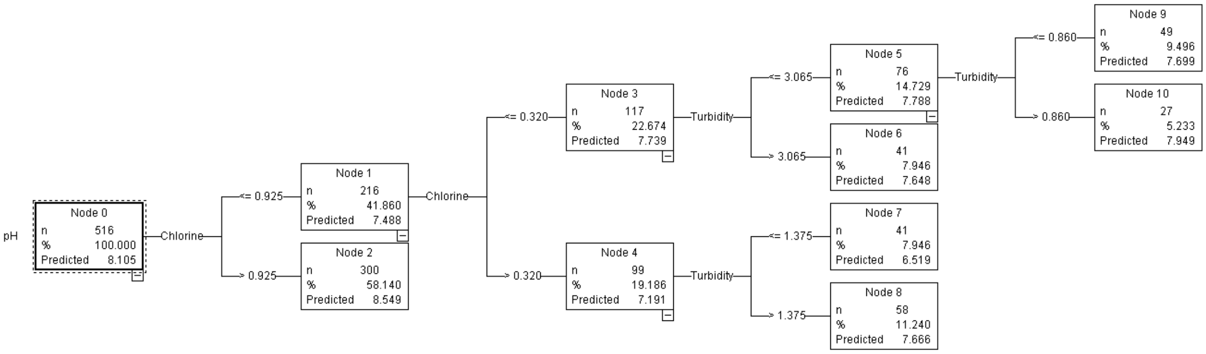

Figure 6. The decision tree in which the result set is pH

As can be seen from Fig 6, it can be stated that pH diverges from the ideal value when chlorine is greater than 0.92.

4. Conclusion

The decision trees built with the Gini algorithm applied to the data set obtained from the water treatment plant demonstrated that the attributes of chlorine, pH and turbidity measured in the water are interrelated.

Having the knowledge of the two attributes, one may have knowledge of the third attribute by the help of decision trees, hence it may be decided whether the water will be treated or not.

Existence of a decision support system will decrease the need for the employees who measure and check the water regularly. According to results, Chlorine has the biggest effect on Turbidity.

With a decision support system to be developed, one can check the values and decide whether the water will be treated or not.

Acknowledgement

We express our gratitude to the Branch Directorate of SCADA at the Department of Drinking Water and Sewer System in the Directorate General of ISU, Turkey and all authorized people for providing us with the parameter values of pH, turbidity and chlorine used in this study, that have been measured in real time and saved in the database of the SCADA System of the Directorate General of ISU, Turkey.

References

[1] Services of Engineering, Production and Trading of Industrial Substance and Material, Proses,

[online] http://www.proses-tim.com/medya/su-kimyasi.pdf (14.01.2013).

[2] Vartiainen, T., Liimatainen, A., Kauranen, P., Hiisvirta, L., 1988. Relations between drinking water mutagenicity and water quality parameters. Chemosphere 17/1, 189-202.

[3] Abdullah, P., Yee, L. F., Ata, S., Abdullah, A., Ishak, B., Abidin, K. N. Z., 2009. The study of interrelationship between raw water quality parameters, chlorine demand and the formation of disinfection by-products. Physics and Chemistry of the Earth Parts A/B/C 34/13-16, 806-811.

[4] Barbeau, B., Desjardins, R., Mysore, C., Prevost, M., 2005. Impacts of water quality on chlorine and chlorine dioxide efficacy in natural waters. Water Research 39/10, 2024-2033.

[5] Hurley, T., Sadiq, R., Mazumder, A., 2012. Adaptation and evaluation of the Canadian Council of Ministers of the Environment Water Quality Index (CCME WQI) for use as an effective tool to characterize drinking source water quality. Water Research 46/11, 3544-3552.

[6] Abu-Sharar T. M., Salameh A. S., 1995. Reductions in hydraulic conductivity and infiltration rate in relation to aggregate stability and irrigation water turbidity. Agricultural Water Management 29/1, 53-62.

[7] Lu, R. S., Lo, S. L., 2002. Diagnosing reservoir water quality using self-organizing maps and fuzzy theory. Water Research 36, 2265-2274.

[8] Astel, A., Tsakovski, S., Barbieri, P., Simeonov, V., 2007. Comparison of self-organizing maps classification approach with cluster and principal components analysis for large environmental data sets. Water Research 41, 4566-4578.

[9] Zhang, R., Zhao, H., Piao, Y., 2011. Applying data mining and hpc for water quality assessment and prediction. In: 3rd International Conference on Advanced Computer Control, Harbin, January 18-20, 2011.

[10] Lu, J., Huang, T., 2009. Data mining on forecast raw water quality from online monitoring station based on decision-making tree. In: 5th International Joint Conference on INC, IMS and IDC, Seoul, August 25-27, 2009.

[11] Yang, Y. J., Haught, R. C., Goodrich, J. A., 2009. Real-time contaminant detection and classification in a drinking water pipe using conventional water quality sensors: Techniques and experimental results. Journal of Environmental Management 90/8, 2494-2506.

[12] Karimipour, F., Delavar, M. R., Kinaie, M., 2005. Water quality management using GIS data mining. Journal of Environmental Informatics 5/2, 61-72.

[13] Hand, D., Mannila, H., Smyth, P., 2001. Principles of data mining. 1st ed., A Bradford Book The MIT Press, London.

[14] Berson,, A., Smith, S., Thearling, K., 2000. Building data mining applications for CRM. McGraw-Hill Professional Publishing, New York, USA.

[15] Chaudhuri, S., 1998. Data Mining and Database Systems : Where is the Intersection?. IEEE Bulletin of the Technical Committee on Data Engineering 21/1, 4 - 8.

[16] Ozekes, S., Camurcu, A. Y., 2002. Classification and prediction in a data mining application. Journal of Marmara for Pure and Applied Sciences 18, 159-174.

[17] Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C. J., 1984. Classification and Regression Trees. Wadsworth, Belmont.