PEGEM EĞİTİM ve ÖĞRETİM DERGİSİ

PEGEM JOURNAL OF EDUCATION and INSTRUCTION

Article Type:
Research Paper

Original Title of Article:
Comparison of decision trees used in data mining

Turkish Title of Article:
Veri madenciliğinde kullanılan karar ağaçlarının karşılaştırılması

Author(s):
Gökhan AKSU, Nuri DOĞAN

For Cite in:
Aksu, G. & Doğan, N. (2019). Comparison of decision trees used in data mining. *Pegem Eğitim ve Öğretim Dergisi, 9*(4), 1183-1208 http://dx.doi.org/10.14527/pegegog.2019.039

Makale Türü:
Özgün Makale

Orijinal Makale Başlığı:
Comparison of decision trees used in data mining

Makalenin Türkçe Başlığı:
Veri madenciliğinde kullanılan karar ağaçlarının karşılaştırılması

Yazar(lar):
Gökhan AKSU, Nuri DOĞAN

Kaynak Gösterimi İçin:
Aksu, G. & Doğan, N. (2019). Comparison of decision trees used in data mining. *Pegem Eğitim ve Öğretim Dergisi, 9*(4), 1183-1208 http://dx.doi.org/10.14527/pegegog.2019.039

PEGEMAKADEMİ

Pegem Eğitim ve Öğretim Dergisi, 9(4), 2019, 1183-1208

www.pegegog.net

Comparison of decision trees used in data mining

Gökhan AKSU [*a], **Nuri DOĞAN** [**b]

[a] Adnan Menderes University, Education Faculty, Aydın/Turkey
[b] Hacettepe University, Education Faculty, Ankara/Turkey

Article Info

DOI: 10.14527/pegegog.2019.039

Article History:
Received 27 Febuary 2019
Revised 25 July 2019
Accepted 28 August 2019
Online 01 October 2019

Keywords:
Data Mining,
Decision Trees,
Classification,
WEKA,
PISA.

Article Type:
Research paper

Abstract

The purpose of this study is to compare decision trees obtained by data mining algorithms used in various areas in recent years according to different criteria. In the study, similar and different aspects of the decision trees obtained by different methods for classifying the students as successful and unsuccessful in terms of science literacy were revealed with the help of 12 independent variables included in the PISA 2015 student survey. Data collected across Turkey, from a total of 5895 students in the age group of 15, were analyzed in Java-based Weka software, which has an open source code. As a result of the analysis, it was found that the most successful algorithms in terms of correct classification rate were respectively Logistic Model, Hoeffding Tree, J.48, REPTree and Random Tree. In addition, regarding the decision trees obtained by different learning algorithms, variables that have been effective in the classification were found to be different. According to the results, it was concluded that independent variables found to be effective in the classification of the students for the decision trees obtained by different algorithms differed from each other and it was suggested to report the finding of more than one algorithm instead of those of only one algorithm.

Veri madenciliğinde kullanılan karar ağaçlarının karşılaştırılması

Makale Bilgisi

DOI: 10.14527/pegegog.2019.039

Makale Geçmişi:
Geliş 27 Aralık 2019
Düzeltme 25 Temmuz 2019
Kabul 28 Ağustos 2019
Çevrimiçi 01 Ekim 2019

Anahtar Kelimeler:
Veri madenciliği,
Karar ağaçları,
Sınıflama,
WEKA,
PISA.

Makale Türü:
Özgün makale

Öz

Bu çalışmanın amacı son yıllarda farklı alanlarda kullanılan veri madenciliği yöntemleri tarafından elde edilen karar ağaçlarının farklı ölçütlere göre karşılaştırılmasıdır. Çalışmada PISA 2015 öğrenci anketinde yer alan 12 bağımsız değişken yardımıyla öğrencileri fen okuryazarlığı bakımından başarılı ve başarısız olarak sınıflama amacıyla farklı yöntemler tarafından elde edilen karar ağaçlarının benzer ve farklı yönleri ortaya çıkarılmıştır. Türkiye örnekleminde 15 yaş grubundaki toplam 5895 öğrenciden elde edilen veriler Java tabanlı ve açık kaynak kodlu WEKA programında analiz edilmiştir. Analiz sonucunda doğru sınıflama oranları bakımından en başarılı yöntemlerin sırasıyla lojistik model, Hoefding tree, J.48, REPTree ve Random tree olduğu belirlenmiştir. Bunun yanında faklı öğrenme yöntemleri tarafından elde edilen karar ağaçlarında sınıflamada etkili olan değişkenlerin farklılık gösterdiği belirlenmiştir. Elde edilen sonuçlara göre farklı yöntemler tarafından elde edilen karar ağaçlarında öğrencileri sınıflamada etkili olan bağımsız değişkenlerin farklılık gösterdiği ve tek bir yöntem yerine birden fazla yönteme ilişkin bulguların rapor edilmesi önerilmiştir.

* **Author:**
** gokhanaksu1983@hotmail.com
Author: nurid@hacettepe.edu.tr

Orcid ID: https://orcid.org/0000-0003-2563-6112
Orcid ID: https://orcid.org/0000-0001-6274-2016

Introduction

In this era of technology that we are experiencing nowadays, the amount of information is constantly increasing every day (Larose, 2005). Therefore, storing the data on hand and obtaining meaningful information from these unprocessed, raw data are quite difficult (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

The process of sorting large amount of data within large databases through computer programs and making predictions for the future using obtained results is called as data mining (DM) (Thuraisingham, 2003). In order to make predictions for the future, one should go back to the past and should see the relevant information and applications related to this topic. Nowadays, there are many algorithms and software developed for this purpose. Thanks to these algorithms and software, the job of the researchers gets quite easy (Imielinski & Mannila, 1996).

DM, which is also called as information discovery, is used to define the process of analyzing the data from a new perspective and outlining useful new information among this data (Sieber, 2008). Data mining aims to discover the information hidden in a large amount of data after a series of operations, which is quite useful for the researchers, thus it is liken to mining which is performed to obtain ore (Aydın, 2007). The algorithms of these systems are usually based on prediction or classification techniques and the objective of the process is developing empirical data's classification charts from the data on hand, which can be used to predict the behavior of unknown objects (Weiss & Kulikowski, 1991). The problem that forms a decision tree is expressed in a recursive way. Decision tree has a structure with the root at the top and then splitting down with the branches and nodes. In other words, decision trees have an inverted tree structure with the root at the top. The nodes, except the root, are called as internal node or tested node. In a decision tree, each node divides the data into two or more subfields according to a certain allocation function of the input property value of the cases in the data set. The tests performed at the nodes make the classification (differentiation) according to possessing a certain property or not whereas for a numeric property the classification is made based on the values of different ranges regarding a certain cut-off score of the property (Maimon & Rokach, 2005). For each sample, a property, which will be assigned to the root node, is set and a branch is built for each possible value of this property, then this operation is consecutively repeated for each branch for the cases that reach the branch. If all cases in a node belong to the same class, the processing of this tree branch (node) would be stopped, since there won't be any differentiation anymore (Tan, Steinbach, & Kumar, 2005).

In order to classify new data using a decision tree, all consecutive internal nodes are visited starting from root node until reaching a leaf node. The tests related to the node are performed in each internal node and they are recorded. The outcome of an internal node's test determines subsequent branch and the next node to visit. The recorded class is the class of the last leaf node. In this way, the combination of all conditions from the root to a leaf constitutes one of the conditions of the class related to the leaf (Rastogi & Shim, 2000).

In a decision tree, each leaf connected to the node indicates a class value. The only remaining thing to do before making a decision is to determine how to divide each property when a series of data from different classes is given. In order to decide which branch will be the best choice, it will be sufficient to check the number of yes and no classes in the leaves. When a single class in the form of Yes or No is formed in the leaf, no further split will be required and the splitting process moving downward will end. In Figure 1, a decision tree is displayed, which the structural presentation of how to make a classification is starting from the root node. Thanks to this tree, the presentation of the information can be made in a briefer and more attractive way (Barros, Carvalho, & Freitas, 2015).

The tree starts with the root node at the origin, given as level 0. The circles on the tree indicate the root and internal nodes whereas the rectangles indicate leaf nodes. In this example, the decision tree was designed for classification, thus leaf nodes include class labels. In Level 1, the cases are divided into two according to the threshold value of X property. If all of the cases go to a class, as can be seen in the left

side of the tree, then no further sub-branches will be built in the tree. If there are two different class values for a property, as can be seen at the second level of Figure 1, the splitting will continue. The branches of a tree are built as the tests continue.

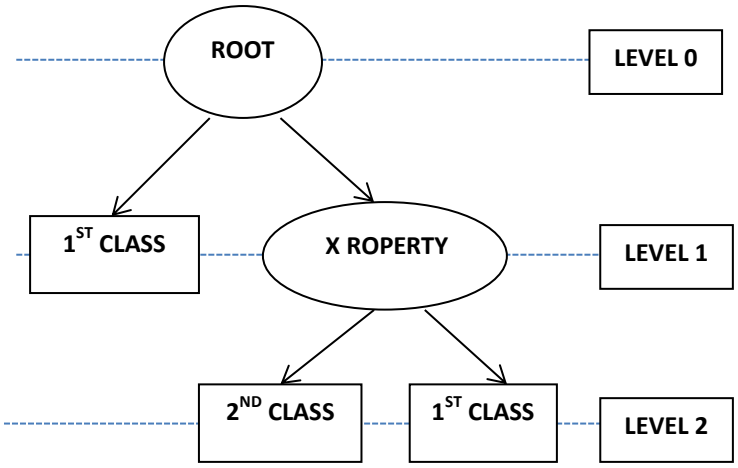


Figure 1. A generic decision tree for classification.

The nodes of a decision tree are for testing a certain property. The tests performed at the nodes are usually in the form of comparing the observed value of a property with a constant value. However, some trees compare two properties with each other or use some functions for one or more properties. Leaf nodes provide a classification or classification cluster for all cases occurring until this leaf is reached and the probability distribution for all classifications. In order to classify an unknown case, it is directed towards the lower branches of the tree according to the values of the properties tested in consecutive nodes; when a leaf is reached, the cases will be classified according to the class assigned to the leaf (Witten & Frank, 2005).

All algorithms and learning methods used in data mining are based on multiple logical test models that have solid statistical basis (Mease & Wyner, 2008). There are three types of data mining techniques, namely unsupervised, semi-supervised and supervised learning approaches. In supervised learning, the algorithm works with a series of cases, whose label is known. These labels may be nominal values for classification tasks as well as numeric values for regression. On the contrary, in unsupervised learning the labels of the cases in the database are unknown and the algorithm typically aims to group the cases according to the similarity of the attribute's values that characterize clustering task. Finally, in semi-supervised learning, a small subset of labelled cases is available, the technique is based on using them along with numerous unlabelled cases (Neelamegam & Ramaraj, 2013). Decision tree is positioned among the basic classifying techniques including the nearest neighbor, support vector machine, Naive Bayes classifier and artificial neural networks, which are all supervised learning approaches (Han & Kamber, 2006). However, clustering analysis algorithms are accepted as unsupervised learning approach because they don't require to set the number of objects in a class or the number of classes (Kusiak, 2001). In addition, Wu et al. (2008) reported that the most popular 10 algorithms in DM are C4.5, k-means algorithm, support vector machine, a priori distribution, Expectation Maximization, PageRank, k-nearest neighbors, Naive Bayes Classifier and Classification and Regression Trees. Among them, C4.5 (Classification Tree) algorithm can be considered as an improved version developed to overcome the limitations of ID3 (Iterative Dichotomiser 3) algorithm (Hssina, Abdelkarim, Ezzikouri, & Erritali, 2014).

The decision tree was first built by Quinlan, under ID3 algorithm in 1986 and then this algorithm was updated and called as C4.5 and it still forms a basis for the decision trees obtained through statistical algorithms. Both ID3 and C4.5 algorithms are based on the statistical computation of the information gain coming from a single attribute. Accordingly, the property that gives maximum information about the decision to be taken based on the cases in the data set is selected and the process is continued by choosing another property that provides maximum information among the remaining ones (Podgorelec, Kokol, Stiglic, & Rozman, 2002). With the recent update, J.48 algorithm replaced them. In short, J.48 algorithm is the decision tree learning algorithm based on the algorithm previously known as C4.5.

In recent years, data mining programs, such as WEKA (Waikato Environment for Knowledge Analysis), Enterprise Mining and Clementine, are successfully applied to the large data sets obtained from geology, medicine, marketing, banking and other commercial areas (Lv, Kim, Zheng, & Jin, 2018). The review of the relevant literature revealed that there is not much application of DM algorithms in education area. Moreover, data mining application were only used for classification purposes in the very few papers and thesis performed in education area. In this study it was aimed to draw meaningful outcomes from the large data of PISA (Program for International Student Assessment), which is one of the large-scale international tests, including a substantial amount of data about the students and schools and to illustrate the use of DM algorithms in education area. In addition, we also aimed to compare the decision trees obtained using different algorithms of classification and prediction techniques, under different conditions. In the literature, there are different algorithms for building a decision tree and it is believed that revealing how they really works with the data obtained from large-scale tests, such as PISA, TIMSS, will be useful. By this means, it will be easier to decide the variables to be preferred in the decision studies to be made in the future. For this purpose, the similarities and differences of the decision trees obtained from PISA 2015 data using different algorithms were determined expecting to form a basis for future studies.

Method

Research Design

In this study, it was intended to predict Science literacy achievement of the students in 15-age group using the answers given to the sub-scales of PISA questionnaire and determine the functioning level of data mining classification techniques that were used in this process. Accordingly, first of all raw data were obtained from PISA questionnaire and then the variables were determined. Afterwards the data to be analyzed were pre-treated and it was analyzed using data mining techniques. Suggestions were submitted according to the findings and results obtained at the end of the analysis.

Research model

In this study, it was intended to predict students' achievement in terms of Science literacy through the answers given to questions included in PISA 2015 questionnaire and examine the decision trees obtained in this process. At the first stage of the study, the accuracy rating of the classification algorithms used to predict students' PISA achievement were determined, then at the second stage the similarities and differences of the decision trees obtained by different algorithms were revealed. Since the study predicts achievement status of the students classified as successful and unsuccessful according to PISA science scores through the sub-tests that measured students' affective properties and sociodemographic indexes, it is a basic research model (Büyüköztürk, Kılıç Çakmak, Akgün, Karadeniz, & Demirel, 2016). Basic researches are defined as studies that produce knowledge and theory and studies based on methodological analyzes are evaluated within this scope (Vaus, 2001). The study is thought to be a basic research since it is aimed to determine the similar and different aspects of the classes obtained by data mining methods and the decision trees used to predict PISA science achievement (Fraenkel, Wallen, & Hyun, 2012).

The Population and Sample of the Research

Regarding the objectives of the research, the study group was formed by the students in the age group of 15 who were registered in formal education and who participated in PISA 2015 test organized by OECD. A total of 540000 students from 72 countries have participated in the test, and 5895 of them were from Turkey. Regarding the execution of PISA 2015 in Turkey, the population of the students in the age group of 15 consisted of 1324089 students; whereas accessible population was 925366 students. In PISA research, school sample was determined according to stratified random sampling method (MoNE, 2016).

Data Collection Process

The data used in the research were acquired from the official web site of OECD, which was opened to the public in 2017, found in the address http://www.oecd.org/pisa/data/2015database/. The data of 5898 students, having country code TR in the student survey stored in SPSS data file format, were used as data source for the analysis.

Data Analysis

The objective of the first stage of the research was to develop a model to predict students' science literacy performance using their PISA data. Since the performance of the students on the PISA test was coded as "Successful" or "Unsuccessful", this was a classification problem and it was analyzed using data mining classification techniques (Romero & Ventura, 2013). Thus, the analyses were performed by using a total of 5 different learning algorithms, namely Hoeffding Tree, J.48, Logistic Model, REPTree and Random Tree that were frequently used in the literature for classification and prediction purposes, allowing the formation of decision tree in Weka software. At the second stage of the study, the variables found to have significant effects on the decision trees obtained through different algorithms were determined and their hit ratio were compared according to their correct classification rate. In addition, the other criteria to be used in the comparison of the models were set as Kappa statistic, mean absolute error, root-mean-square error, relative absolute error and route relative squared error. Since there is no standard criterion for comparing these values, the errors should be as low as possible whereas Kappa statistic and correct classification rate should be as high as possible (Almuniri & Said, 2017; Doreswamy, 2012; Hossin & Sulaiman, 2015; Kiranmai & Damodaram, 2014; Sokolova & Lapalme, 2009; Tiwari, Jha, & Yadav, 2012; Vihinen, 2012).

Kappa statistic or Kappa value is a numerical value comparing expected and observed values and is considered less misleading since it also takes into account the chance factor (Witten & Frank, 2005). The accuracy rate observed in this process is obtained by dividing the number of samples correctly classified over the entire matrix by the total number of samples. Expected accuracy means that any random classifier is successful based on the given matrix. Finally, after determining the expected and observed accuracy rates, Kappa statistics are calculated with the help of the equation given below (Carletta, 1993).

$$\text{Kappa } (\kappa) = \frac{\text{Observed accuracy value} - \text{Expected accuracy value}}{1 - \text{Expected accuracy value}}$$

Thus, Kappa statistic checks the expected accuracy of a random classifier by the machine learning method and gives a measure of how close to realistic the samples in the data set are. Even though there is not an absolute standard in the interpretation of Kappa statistic, usually the values between .00-.20 are defined as none to slight, .21-.40 as fair, .41-.60 as moderate; .61-.80 as substantial and .81-1.00 as almost perfect (Landis & Koch, 1977).

The Software and Algorithms Used in the Data Analysis

In the analysis of data, java based WEKA 3.8 (Waikato Environment for Knowledge Analysis) software was used. WEKA software was developed by Waikato University of New Zeeland to process agricultural data (Kuyucu, 2012). The main reasons for preferring WEKA software in the data analysis are the widespread use of the software and its open source system.

ID3, which was one of the most popular decision trees, was a stable algorithm founded by J.Ross Quinlen from Austria Sidney University, based on information gain criterion. Afterward, C4.5 tree algorithm, which possesses many novelties and improvements such as allowing dependent variable to be numeric, missing data, noisy data and setting rule from the trees, was developed on the basis of ID3.

Logistic Model Trees (LMT) may work with dual-category or multiple-category properties, numeric properties or classification level properties of the property defined for the target (result) variable, and with missing data. When the logistic regression function is executed in a node, cross validation algorithm is used to determine the number of iterations needed for the analysis, thus the same number is used across the whole tree instead of performing a different iteration in each node. This operation considerably increases the running time while having very few impacts on the accuracy of the obtained results. As an alternative, you can set the number of iterations to be used across the tree by yourself. Normally, decreasing cross validation number is an incorrect classification error. Instead of this, root-mean-square error of the probabilities may be chosen as well. The division criterion of the tree is determined according to the amount of information that C4.5 has or logit residual values. Here, the efforts are made to increase the purity of residual values.

Regression Tree (REPTree) builds a decision/regression tree using information gain and variance reduction. Variance reduction is performed using reduced-error pruning. This algorithm deals with missing values by splitting the corresponding cases into pieces. The minimum number of cases per leaf can be set by the users. For maximum tree depth, the minimum numeric class variance, proportion of train variance for split and the number of levels for splitting can be chosen by making changes in the window opened by ticking the command line next to the selection button.

In Random Tree algorithm, a pre-defined number of attributes are randomly selected and the test is conducted without any pruning at the branches of the tree. In decision trees, the best functioning property, in other words the property that provides maximum information is selected; however, in Random Tree algorithm this selection is made randomly (Witten, Frank, & Hall, 2016).

Results

As a result of the relevant literature review, the number of variables to be used for predicting PISA science literacy was found to be 29; thus, first of all the best properties were determined using different algorithms. The information about the variables found to be effective on the science literacy of the students and used within the scope of the study are displayed in Table 1, with their name, code, minimum and maximum values, as well as their percentile values about the number of layers that each variable was successful in 10-level cross validation algorithm. Even though Artificial Neural Networks algorithms were reported to make better predictions compared to regression analysis in case of a large number of variables added to the model (Lykourentzou, Giannoukos, Mpardis, Nikolopoulos, & Loumos, 2009); it was also reported that in predictive algorithms such as DM, adding more variables to the model will not lead to an increase on the accuracy of the performance estimates (Huang & Fang, 2013) . In addition, in the experimental study performed by (Kohavi, 1995) it was reported that the estimates found under different algorithms were quite good and almost unbiased when the number of layers was 10 and 20; thus 10-level cross validation algorithm was used in the study. At the end of this process, the stage of comparing data mining algorithms used for predicting PISA science literacy through 12 independent variables, according to correct classification numbers, correct classification rate, reliability values and decision trees was started. At this stage, PISA science literacy variable was converted into a categorical variable by WEKA software. Students who got a score lower than 425.00, which is Turkey's average, were coded as unsuccessful (0) whereas those who got a higher score were coded as successful (1).

Table 1.
Summary Information for the Variables that were Used in Data Mining.

No	Name of the variable	Code	min	max	Hit Ratio
1	Inquiry-based science education	IBTEACH	-3.34	3.18	%100.00
2	Environmental awareness	ENVAWARE	-3.38	3.29	%100.00
3	Scientific beliefs	EPIST	-2.79	2.16	%100.00
4	Vocational status expected from the student	BSMJ	10.00	89.00	%100.00
5	Duration of studying outside of school (week)	OUTHOURS	.00	70.00	%100.00
6	Time of studying science (week)	SMINS	.00	800.00	%100.00
7	Overall studying time (week)	TMINS	100.00	3000.00	%100.00
8	Test anxiety	ANXTEST	-2.51	2.55	%60.00
9	Fairness of the teacher	Unfairteacher	1.00	24.00	%40.00
10	Educational resources at home	HEDRES	-4.37	1.18	%80.00
11	ICT resources	ICTRES	-3.27	3.50	%100.00
12	Socio-economic status index	ESCS	-5.13	3.12	%100.00

Findings about Different Learning Algorithms

The results obtained using 10-level cross validation of the measurement outcomes obtained by Hoeffding Tree, J.48, Logistic Model, REPTree and Random Tree algorithms for predicting the achievement using the variables measured by PISA questionnaire are respectively reported below.

Findings of Hoeffding Tree Algorithm

4085 of 5865 students were correctly classified using Hoeffding Tree algorithm, resulting with 69.65% correct classification rate. Kappa statistic of the classification was found to be .39 and mean absolute error was .36. The tree structure built by the software for Hoeffding tree algorithm is shown in Figure 2.

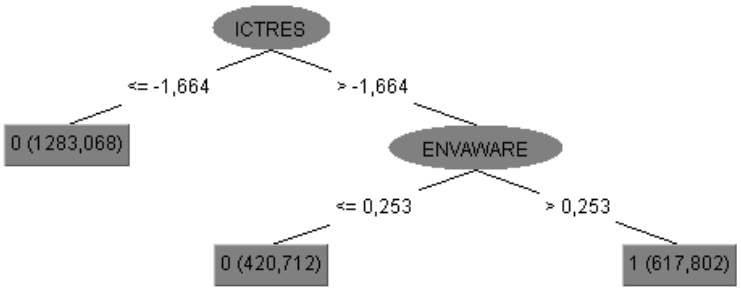


Figure 2. Tree structure obtained by Hoeffding tree.

It was found that Information and Communication Technology resources was the most effective variable for classifying students in terms of PISA science literacy, followed by environmental awareness. The tree structure was built by the program, no further split occurred at lower levels. In other words, decision tree was set as two levels. Regarding the decision tree, it can be seen that the cut-off score of Information and Communication Technology resources (ICTRES) was -1.64, those who were under this value were classified as unsuccessful whereas for those who were above this value environmental awareness score, whose cut-off score was .25, was checked. The measure of the purity used in the classification is called as information and measured by the units defined as ICT. Taking the amount of information, which is called as entropy, as the basis, it was concluded that only 2 of the 12 variables covered within the scope of the research were sufficient for classifying students as successful and unsuccessful.

Findings of J.48 Algorithm

3976 of 5865 students were correctly classified using J.48 algorithm, resulting with 67.79% correct classification rate. Kappa statistic of the classification was found to be .35 and mean absolute error was .38. A portion of the tree structure built by the software for J.48 algorithm is shown in Figure 3.

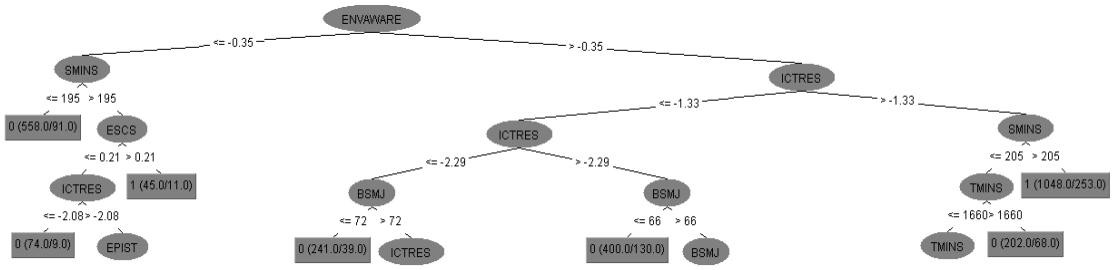


Figure 3. Tree structure obtained by J.48.

The decision tree obtained via J.48 algorithm has a total of 15 levels but for the sake of ease of interpretation only the first four levels are displayed. It was found that environmental awareness was the most effective variable for classifying students in terms of PISA science literacy, followed by weekly time of studying science and ICT resources. Regarding the next level of classification, it can be seen that the effective variables were socio-economic status index, ICT resources and weekly time of studying science. The cut-off score of environmental awareness was calculated from the raw scores by the software and it was found as -.35. Those who were under this value were checked in terms of weekly time of studying science whereas for those who were above this value, ICT resources were checked. The cut-off score of weekly time of studying science was 195.00. Then, socio-economic status index was checked for those whose time of studying science was above 195.00 whereas those whose time of studying science was below 195.00 classified as unsuccessful. The cut-off score of socio-economic status index, which was located at the left branch of the decision tree, was found to be .21 and those who were above this value were classified as successful whereas ICT resources was checked for those who were under this value.

Findings of Logistic Model Algorithm

4125 of 5865 students were correctly classified using Logistic Model (LM), resulting with 70.33% correct classification rate. Kappa statistic of the classification was found to be .40 and mean absolute error was .38. The tree structure built by the software for J.48 algorithm is shown in Figure 4.

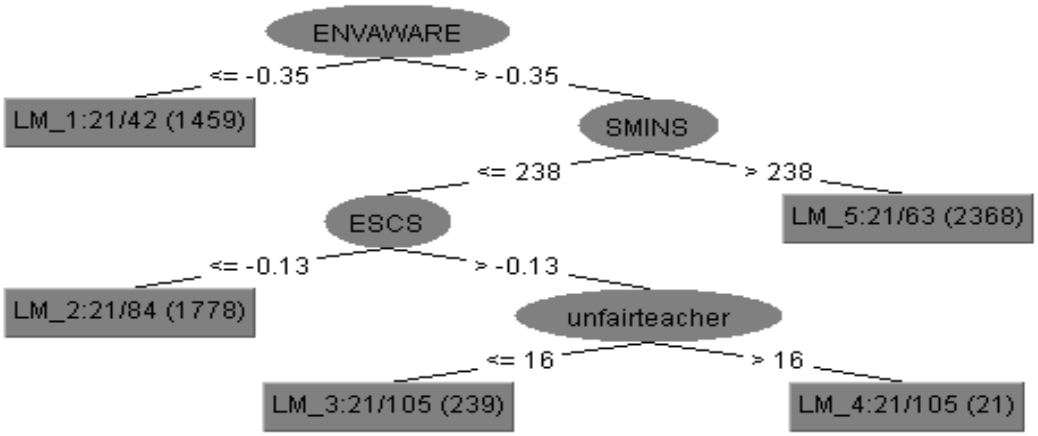


Figure 4. Tree structure obtained by Logistic Model.

The obtained tree structure was built by the software and no further split occurred at lower levels. In other words, the decision tree was found to be with three levels. It was found that environmental awareness was the most effective variable for classifying students in terms of PISA science literacy, followed by weekly time of studying science. It can be seen that and socio-economic status index variable was effective in the next level of classification whereas fairness of the teacher variable was effective in the last level. The cut-off score of environmental awareness was -.35 and those who were under this value were classified according to Logistic Model number 1 whereas weekly time of studying science, whose cut-off score was 238.00, was checked for those who were above this value. It was found that socio-economic status index was effective on classifying students whose time of studying science was below 238.00 hours whereas those whose time of studying science was above 238.00 hours were directly classified according to the model called LM_5. From Figure 4, it can be seen that the equations formed on the leaves of the decision tree were based on five different logistic model, namely LM_1, LM_2, LM_3, LM_4 and LM_5. In traditional statistical analysis, a single regression model was built for predicting dependent variable using independent variables; however, five different regression models were obtained in Logistic Model. Thus, the Logistic Model's hit ratio is higher than the other algorithms. In addition, it was found that the constants and coefficients of the variables differ in each equation.

Findings of REPTree Algorithm

3946 of 5865 students were correctly classified using REPTree, resulting with 67.28% correct classification rate. Kappa statistic of the classification was found to be .34 and mean absolute error was .39. A portion of the tree structure built by the software for REPTree algorithm is shown in Figure 5.

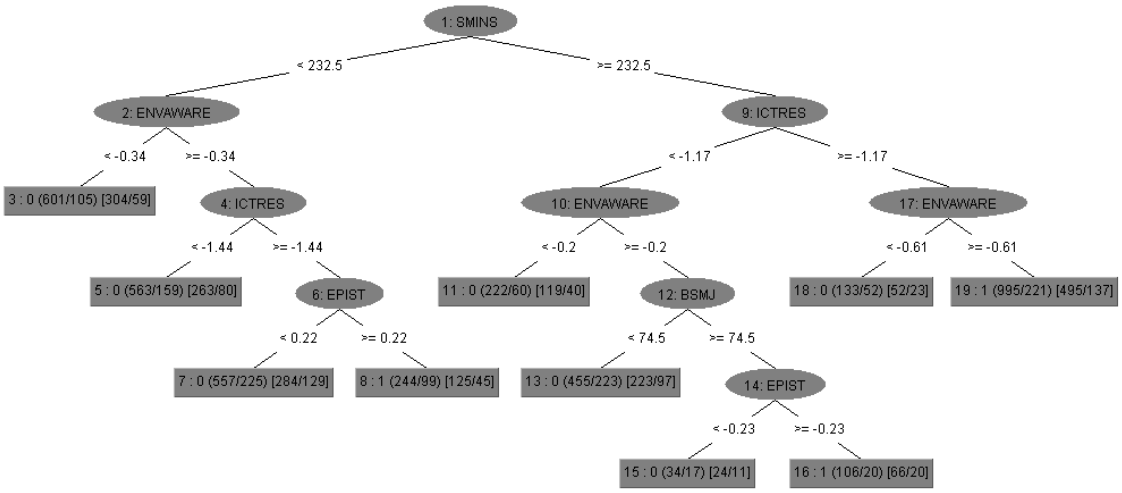


Figure 5. *Tree structure obtained by REPTree.*

The decision tree obtained via REPTree algorithm has a total of 13 levels but for the sake of ease of interpretation only the first three levels are displayed. It was found that weekly time of studying science was the most effective variable for classifying students in terms of PISA science literacy, followed by environmental awareness and ICT resources. Regarding the next level of classification, it can be seen that environmental awareness variable was effective. In the first classification of the tree, the cut-off score of weekly time of studying science was found to be 232.50 and environmental awareness score of those who were under this value was checked whereas ICT resources score was checked for those who were above this value. Similarly, cut-off scores of ICT resources on both branches were found to be -1.67 and -1.44 respectively.

Findings of Random Tree Algorithm

3690 of 5865 students were correctly classified using Random Tree, resulting with 62.92% correct classification rate. Kappa statistic of the classification was found to be .26 and mean absolute error was .37. A portion of the tree structure built by the software for Random Tree algorithm is shown in Figure 6.

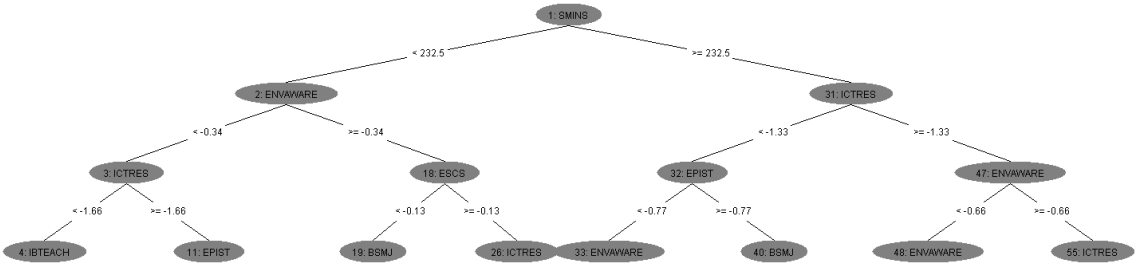


Figure 6. Tree structure obtained by Random Tree.

The decision tree obtained via Random Tree algorithm has a total of 24 levels but for the sake of ease of interpretation only the first three levels are displayed. It was found that weekly time of studying science was the most effective variable for classifying students in terms of PISA science literacy, followed by environmental awareness and ICT resources. Regarding the next level of classification, it can be seen that ICT resources, scientific beliefs, environmental awareness and socio-economic status index variables were effective. In the first classification of the tree, the cut-off score of weekly time of studying science was found to be 232.50 and environmental awareness score of those who were under this value was checked whereas ICT resources score was checked for those who were above this value. The cut-off score of environmental awareness was set as -.34; ICT resources were effective on those who are below this value whereas socio-economic status index was effective on those who are above this value.

Comparison of Different Techniques

In order to compare the reliability values obtained from different algorithms, weighted overall average values were compared. Number of correct classifications, correct classification rate, Kappa statistic, Kappa level and mean absolute error of each algorithm are displayed in Table 2.

Table 2.
Reliability Criteria Obtained by Using Different Algorithms.

Algorithms	Number of levels used in the classification	Number of correct classifications	Correct classification rate	Kappa statistic	Kappa level	Mean absolute error
1.Hoeffding Tree	2	4085	69.65	.39	Fair	.36
2. J.48	15	3976	67.79	.35	Fair	.38
3. Logistic Model	3	4125	70.33	.40	Moderate	.38
4. REPTree	13	3946	67.28	.34	Fair	.39
5. Random Tree	24	3690	62.92	.26	Fair	.37

Regarding Table 2, the best results were achieved by Logistic Model technique in terms of number of correct classifications, correct classification rate, Kappa statistic and Kappa level whereas the best results in terms of mean absolute error were achieved by Hoeffding Tree algorithm. If the ranking was done according to Kappa statistic, the ranking would be as Logistic Model, Hoeffding tree, J.48, REPTree and Random Tree algorithms. Accordingly, it can be seen that ranking the algorithms according to correct classification or Kappa statistic would not create any difference.

Discussion, Conclusion & Implementation

Since the number of variables to be used for predicting PISA science literacy was found to be too much, first of all the properties to be used within the scope of the study were determined. In order to decrease the number of variables, some DM techniques, namely Best First Forward, Best First Backward and Greedy Stepwise were used for 10-level cross validation and PISA science literacy was predicted using 12 variables that had 40.00% or above hit ratio. It was found that the number of levels belonging to the decision trees obtained from different algorithms had no direct impact on the correct classification rate.

Bakker (2016) stated that the decision trees built from the same data set but using different algorithms may differ. Hence, within the scope of the study, the variable that best predicted science literacy was found to be time of studying science for REPTree and Random Tree whereas it was environmental awareness for Logistic Model, and Information and Communication Technology resources for Hoeffding tree. Although the property that gave maximum information was found to be time of studying science for both REPTree and Random Tree algorithms, the next level of the decision trees built by these two algorithms differed from each other; environmental awareness and ICT resources were found to be effective for REPTree algorithm whereas environmental awareness and educational resources at home were effective for Random Tree algorithm. Similarly, it is noted that there were differences in the decision tree obtained by the other algorithms. This result bears a resemblance to the finding of (Mohan, 2013), indicating that when you change the parameters of a technique in itself, no differences would occur in the criteria such as precision, sensitivity, and recovery, however there will be differences among the results obtained by using different techniques.

Regarding the comparison of correct classification rate obtained from different decision tree techniques, it was found that they were ranked as Logistic Model, Hoeffding tree, J.48, REPTree and Random Tree. Accordingly, correct classification rates differ according to the algorithm used for the decision tree. This result shows similarity with the study of Cinaroglu (2016), reporting that there were significant differences among the hit ratios of C4.5, CART and Random forest algorithms. In a similar study, Dietterich (2000) stated that bagging, boosting and randomization methods would create differences on the obtained results. From this aspect, the findings of the study show similarity.

In the study, the best prediction algorithm was found to be Logistic Model according to some reliability criteria, namely number of correct classifications, Kappa statistic and mean absolute error. This result is in line with the studies of Liaw and Wiener (2002) and Svetnik, Liaw, Tong and Wang (2004). It is believed that the reason of getting more reliable results with less error in the decision tree built by Logistic Model is using multiple regression equations instead of one and building learning algorithms that classify new data using the results obtained from the predictions of the regression equations (Strobl, Malley, & Tutz, 2009). It was found that according to the number of error and reliability, the algorithms were ranked as Hoeffding Tree, J4.8, REPTree and finally Random Tree. J 4.8 algorithm, which was considered to be the most widely used algorithm in the literature showed a moderate level of success. On the other hand, Logistic Model and Hoeffding tree algorithms were found to be more successful in terms of both error-based reliability values and validity criteria.

Suggestions

In the study, it was found that input variables that affect output variable differed for the decision trees built by different techniques. Based on these findings, regarding the studies concerning the prediction of academic achievement, it was suggested to provide evidence about the validity of obtained results by using at least 3 learning techniques in data mining instead of using only one.

It was found that Logistic Model (LM) and Hoeffding tree techniques used in data mining, especially in Weka software, provided more reliable and valid outcomes under different conditions, thus it was suggested to use these methods in the studies performed for classification or prediction of the achievement purposes. Especially, the logistic equations obtained from the leaves of the decision tree

built for determining successful and unsuccessful students may help to improve correct classification rate allowing the use of many regression equations for different attribute values instead of modelling the relationships between variables with a single regression equation.

Since the analysis was performed using PISA data, it can be suggested to analyze the result obtained from large-scale tests organized by OSYM using the techniques based on data mining instead of standard statistical techniques based on correlation and regression for the purpose of providing richer evidences.

.

Türkçe Sürüm

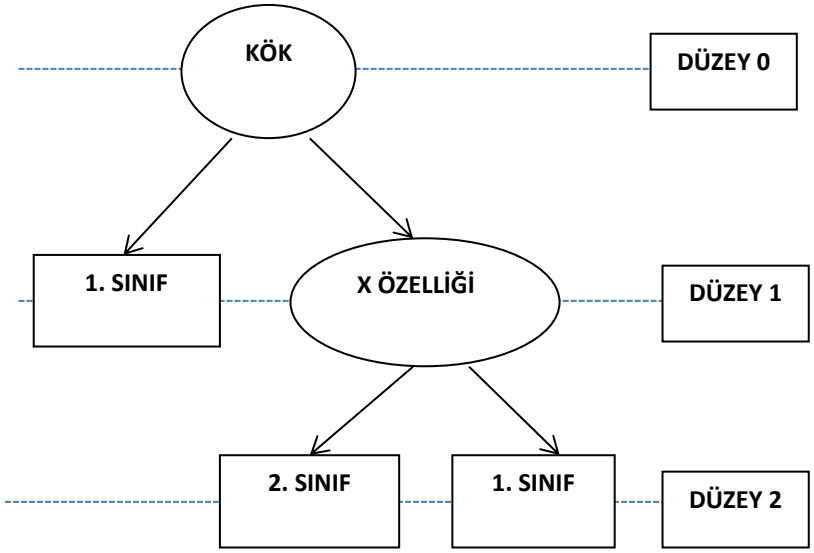
Giriş

Teknoloji çağının yaşandığı günümüzde her geçen gün bilgi miktarı sürekli artış göstermektedir (Larose, 2005). Bu sebeple sahip olunan bilginin depolanması ve bu işlenmemiş ham verilerden anlamlı bilgiler elde etmek oldukça zorlaşmaktadır (Fayyad et al., 1996). Büyük miktarlardaki verilerin büyük veri tabanlarında bilgisayar programları vasıtasıyla aranarak, elde edilen sonuçlar yardımıyla gelecekle ilgili tahmin yapılması işlemlerine veri madenciliği (VM) denilmektedir (Thuraisingham, 2003). Geleceğe dair tahmin yapılabilmesi için geçmişe dönüp, geçmişte bu konularla ilgili ne gibi bilgiler olduğunu ve ne gibi uygulamalar yapıldığını görmek gerekir. Günümüzde bu amaçla birçok algoritma ve yazılım geliştirilmiştir. Bu algoritma ve yazılımlar sayesinde, araştırmacıların işleri oldukça kolaylaşmıştır (Imielinski & Mannila, 1996).

Bazen bilgi keşfi olarak adlandırılan VM veriyi yeni bir perspektiften analiz etmek ve bu veriler arasındaki yararlı yeni bilgileri özetleme sürecini tanımlamak için kullanılmaktadır (Sieber, 2008). VM'de büyük miktardaki verinin içinde saklı olan ve araştırmacılar için oldukça faydalı olan bilgilerin bir dizi işlemlerin ardından ortaya çıkarılması hedeflendiğinden cevher elde etmek için yapılan madenciliğe benzetilmektedir (Aydın, 2007). Bu sistemlerin algoritmaları genel olarak tahminleme veya sınıflama teknikleri üzerine kuruludur ve eldeki verilerden henüz bilinmeyen nesnelerin davranışlarını tahmin etmek için kullanılabilen ampirik verilerin sınıflandırma şemalarının oluşturulmasını hedeflemektedir (Weiss & Kulikowski, 1991). Bir karar ağacı oluşturma problemi kendi kendini tekrarlar şekilde ifade edilmektedir. Karar ağacı en üstte kökü ve daha sonra aşağı doğru dal ve düğümleri olan bir yapıya sahiptir. Başka bir ifadeyle karar ağaçları kökü yukarda olan ters bir ağaç yapısına sahiptir. Kök haricindeki diğer düğümlere içsel düğüm veya test edilen düğüm adı verilmektedir. Bir karar ağacında her bir düğüm, veri setindeki örneklerin girdi özellik değerlerinin belli bir ayırma fonksiyonuna göre iki veya daha fazla alt alana bölmektedir. Düğümlerde gerçekleştirilen testlerde tek bir özelliğe sahip olup olmama durumuna göre sınıflama (ayrışma) yapılırken sayısal bir özellik için özelliğin belli bir kesme puanına göre farklı aralıklardaki değerleri esas alınarak sınıflama yapılır (Maimon & Rokach, 2005). Her bir örnek için kök düğüme yerleşecek şekilde bir özellik belirlendikten ve olası her bir değer için bir şube (dal) oluşturulduktan sonra bu işlem her bir dala ulaşan örnekleri kullanarak her bir dal için ardışık olarak tekrar edilir. Eğer herhangi bir durumda bir düğümdeki tüm örnekler aynı sınıfa dahil olurlarsa ağacın o parçasının (düğümü) geliştirmesi durdurulur. Çünkü artık farklı sınıflara ayrışma olmayacaktır (Tan et al., 2005).

Bir karar ağacı kullanarak yeni verileri sınıflandırmak için, kök düğümden başlayarak, bir yaprağa ulaşılana kadar ardışık iç düğümler ziyaret edilir. Her iç düğümde, düğüme ilişkin testler yapılarak bunlar kayıt altına alınır. Bir iç düğümdeki testin sonucu, geçilen dalı ve ziyaret edilecek sonraki düğümü belirler. Kayıt altına alınan sınıf sadece son yaprak düğümünün sınıfıdır. Böylece, kökten bir yaprağa kadar olan dallar için tüm koşulların birleşimi, yaprak ile ilişkili sınıfın koşullarından birini oluşturur (Rastogi & Shim, 2000).

Karar ağacında düğüme bağlı olan her bir yaprak bir sınıf değerini göstermektedir. Karar vermek için geriye kalan tek şey, farklı sınıflarda bir dizi örnek verildiğinde, hangi özelliğin nasıl bölüneceğini belirlemektedir. Elde edilen dallanmalardan hangisinin en iyi seçim olacağına karar vermek için yapraklardaki evet ve hayır sınıflarının sayılarına bakmak yeterli olacaktır. Sadece Evet veya Hayır şeklinde yaprakta tek bir sınıf oluştuğunda tekrardan ayrılmaya gerek kalmayacak ve aşağıya doğru yinelenen dallanma süreci sona erecektir. Şekil 1'de belirlenen bir özelliğe göre kök düğümden (node) başlayarak nasıl bir sınıflama yapılacağı ile ilgili yapısal bir gösterim olan karar ağacı verilmektedir. Bu ağaç sayesinde daha özet ve daha çekici bilgi sunumu yapılabilmektedir (Barros et al., 2015).



Şekil 1. Sınıflama için genel bir karar ağacı örneği.

Ağaçta ilk olarak düzey 0 olarak verilen başlangıç noktasında kök düğüm ile başlanmaktadır. Ağaçta yer alan çemberler kök ve iç düğümleri, kareler ise yaprak düğümlerini belirtir. Bu özel örnekte, karar ağacı sınıflandırma için tasarlanmıştır ve bu nedenle yaprak düğümleri sınıf etiketleri barındırmaktadır. Düzey 1'de herhangi bir X özelliği için bu özelliğin belli bir eşik değerine göre örnekler iki sınıfa ayrılmaktadır. Ağacın sol tarafında görüldüğü gibi tüm örnekler tek bir sınıfa gidiyorsa bu durumda ağacın alt dalları oluşmayacaktır. Eğer Şekil 1'de ikinci düzeyde görüldüğü gibi belirli bir özellik için iki farklı sınıf değeri varsa dallanma devam edecektir. Bu şekilde testler devam ettikçe ağacın dalları oluşmaktadır.

Karar ağacında yer alan düğümler belirli bir özelliği test etmeyi amaçlamaktadır. Genellikle düğümler üzerinde gerçekleştirilen testler, bir özelliğe ilişkin gözlenen değerin sabit bir değer ile karşılaştırması şeklinde yapılmaktadır. Buna rağmen bazı ağaçlar iki özelliği birbiriyle karşılaştırır veya bir ya da daha fazla özellik için bazı fonksiyonlar kullanırlar. Yaprak düğümler o yaprağa ulaşıncaya kadarki tüm örnek durumlar için bir sınıflama veya sınıflama kümesi ve olası tüm sınıflamalara ilişkin olasılık dağılımlarını vermektedir. Bilinmeyen bir örneği sınıflamak için, ardışık düğümlerde test edilen özelliklerin değerlerine göre ağacın alt dallarına yönlendirme yapılır ve bir yaprağa ulaşıldığında, örnekler yaprağa atanan sınıfa göre sınıflandırılmış olacaktır (Witten & Frank, 2005).

Veri madenciliğinde kullanılan algoritma ve öğrenme yöntemlerinin tamamı sağlam istatistiksel temelleri olan çoklu mantıksal sınama modellerine dayanmaktadır (Mease & Wyner, 2008). Veri madenciliği yöntemlerinin denetimsiz (unsupervised), yarı denetimli (semi-supervised) ve denetimli (supervised) öğrenme yaklaşımı olmak üzere üç türü bulunmaktadır.

Denetimli öğrenmede algoritma, etiketleri bilinen bir dizi örnekle çalışır. Etiketler, sınıflandırma görevi için nominal değerler veya regresyon durumunda ise sayısal değerler olabilir. Denetimsiz öğrenmede, aksine, veri kümesindeki örneklerin etiketleri bilinmemektedir ve algoritma, tipik olarak örnekleri, kümeleme görevini karakterize eden öznitelik değerlerinin benzerliğine göre gruplandırmayı amaçlamaktadır. Son olarak, yarı denetimli öğrenmede ise etiketli örneklerin küçük bir alt kümesi mevcutken bunları çok sayıda etiketlenmemiş örnekle birlikte kullanmayı esas almaktadır (Neelamegam & Ramaraj, 2013). Karar ağacı, en yakın komşuluk, destek vektör makinesi, Naive Bayes sınıflandırıcısı ve yapay sinir ağları temel sınıflandırma yöntemleri arasında yer almaktadır ve bunlar denetimli öğrenme yaklaşımlarıdır (Han & Kamber, 2006). Ancak kümeleme analizi algoritmaları, bir sınıftaki nesne sayısının veya sınıf sayısının belirtilmesini gerektirmediği için denetimsiz öğrenme yaklaşımı olarak kabul edilmektedir (Kusiak, 2001). Bunun yanında Wu et al. (2008) VM'de en popüler 10 algortimanın C4.5, k-ortalama yöntemi, destek vektör makineleri, önsel dağılımlar (apriori), maksimum beklenti

(EM=Expectation Maximization), sayfa derecesi (PR=PageRank), k-en yakın komşuluk (kNN=k-nearest neighbors), Naive Bayes ile Sınıflama ve Resgresyon Ağacı (CART= Classification and Regression Trees) olduğunu belirtmektedir. Bu yöntemlerden C4.5 (Classification Tree) algortiması ID3 (Iterative Dichotomiser 3) yönteminin sınırlılıklarını ortadan kaldırmak için geliştirilmiş bir versiyonu olarak düşünülebilir (Hssina et al., 2014).

İlk olarak 1986 yılında Quinlan tarafından ID3 algoritması altında karar ağacı oluşturulmuş ve sonrasında bu algoritma güncellenerek C4.5 adını almıştır ve halen geleneksel istatistiksel yöntemlerle elde edilen karar ağaçlarına temel teşkil etmektedir. ID3 ve C4.5 algoritmalarının her biri karar ağacı oluşturmak için tek bir öznitelikten gelen bilgi kazancının istatistiksel olarak hesaplanmasına dayalıdır. Bu şekilde eğitim veri setindeki örneklere dayalı olarak alınacak kararla ilgili en fazla bilgiyi veren bir özellik seçilir ve kalan özelliklerden en fazla bilgi veren bir başka özellik daha seçilerek işleme devam edilir (Podgorelec et al., 2002). Son yıllarda yapılan güncelleme ile bunların yerine J.48 algoritması getirilmiştir. Özetle J.48 algoritması daha önce C4.5 olarak bilinen algoritmaya dayalı karar ağacı öğrenme yöntemidir.

Son yıllarda WEKA (Waikato Environment for Knowledge Analysis), Enterprise Mining ve Clementine gibi veri madenciliği programları, jeoloji, tıp, pazarlama, bankacılık ve diğer ticari alanlarda elde edilen büyük veri gruplarına başarılı bir şekilde uygulanmaktadır (Lv et al., 2018). İlgili alan yazın incelendiğinde VM yöntemlerinin eğitim alanında çok fazla uygulaması olmadığı görülmektedir. Bunun yanında eğitim alanında çok az sayıda yapılan makale ve tezlerde ise veri madenciliği uygulamalarının sadece sınıflama amaçlı kullanıldığı belirlenmiştir. Bu çalışmada uluslararası geniş ölçekli sınavlardan biri olan PISA (Programme for International Student Assessment) sınavı ile öğrenci ve okula ilişkin oldukça fazla verinin içerisinden anlamlı sonuçlar çıkarmak ve eğitim alanında VM yöntemlerinin kullanılmasını örneklemek amaçlanmaktadır. Bunun yanında sınıflama ve tahmin yöntemlerinin farklı algoritmaları kullanılarak elde edilen karar ağaçlarının farklı koşullar altında karşılaştırılması amaçlanmaktadır. Alanyazında farklı karar ağacı belirleme yöntemleri bulunmakta ve bunların PISA, TIMSS gibi geniş ölçekli sınavlardan elde edilen verilerde nasıl çalıştığının belirlenmesinin yararlı olacağı düşünülmektedir. Bu sayede ileride yapılacak karar çalışmalarında hangi değişkenlerin tercih edilmesi gerektiğine ilişkin karar vermek daha kolay olacaktır. Bu sebeple araştırmada PISA 2015 verileri yardımıyla farklı algoritmalardan elde edilen karar ağaçlarının benzer ve farklı yönleri belirlenerek ilerde yapılacak çalışmalara temel oluşturması amaçlanmıştır

Yöntem

Araştırmanın Deseni

Çalışmada ülkemizde 15 yaş grubundaki öğrencilerin PISA öğrenci anketinde yer alan alt ölçeklere verdikleri yanıtlar yardımıyla Fen okuryazarlığı başarılarını tahmin etmek ve bu süreçte kullanılan veri madenciliği sınıflama yöntemlerinin ne düzeyde işlediğini belirlemek amaçlanmıştır. Buna göre öncelikle PISA öğrenci anketinden ham veriler elde edilmiş ve ardından değişkenler belirlenmiştir. Sonrasında analiz edilecek veriler ön işleme tabi tutulmuş ve veri madenciliği yöntemleri ile analiz edilmiştir. Analiz sonucunda elde edilen bulgular ve sonuçlara dayalı olarak önerilerde bulunulmuştur.

Araştırmanın Modeli

Bu araştırmada öğrencilerin PISA 2015 öğrenci anketinde yer alan sorulara verdikleri yanıtlar yardımıyla Fen okuryazarlığı bakımından başarılarını tahmin etmek ve bu süreçte elde edilen karar ağaçlarının incelemesi amaçlanmıştır. Çalışmanın ilk aşamasında öğrencilerin PISA başarılarını tahmin etmede kullanılan sınıflama yöntemlerinin doğruluk derecesi belirlenmiş ardından ikinci aşamada farklı algoritmalar yardımıyla elde edilen karar ağaçlarının benzer ve farklı yönleri ortaya çıkarılmıştır. Araştırma, PISA fen puanları açısından başarılı ve başarısız olarak sınıflanan öğrencilerin duyuşsal özellikleri ölçen alt testler ve sosyodemografik indeksler yardımıyla başarı durumlarına ilişkin tahminleme yapılması bakımından temel araştırma modelindedir (Büyüköztürk et al., 2016). Temel araştırmalar bilgi ve kuram üretmeye dönük çalışmalar olarak tanımlanmakta ve yöntemsel analizlere dayalı çalışmalar da bu kapsamda değerlendirilmektedir (Vaus, 2001). Çalışmada veri madenciliği

yöntemleriyle elde edilen sınıfların ve PISA fen başarını tahmin etmede kullanılan karar ağaçlarının benzer ve farklı yönlerinin belirlenmesi amaçlandığından çalışmanın temel araştırma niteliğinde olduğu düşünülmektedir (Fraenkel et al., 2012).

Araştırmanın Evreni ve Örneklemi

Araştırmanın amaçları doğrultusunda çalışma grubunu OECD tarafından düzenlenen PISA 2015 sınavına katılan ve örgün eğitime kayıtlı olan 15 yaş grubu öğrencileri oluşturmaktadır. Sınava 72 ülkeden toplam 540000'e yakın öğrenci katılmış ve bunların 5895 tanesi Türkiye'den katılan öğrencilerdir. PISA 2015 Türkiye uygulamasında 15 yaş grubu öğrenci evreni 1324089 öğrenci, uygulamaya katılabilecek ulaşılabilir Türkiye evreni ise 925366 öğrenci olarak belirlenmiştir. PISA araştırmasında okul örneklemi, tabakalı seçkisiz örnekleme yöntemiyle belirlenmektedir (MEB, 2016).

Veri Toplama Süreci

Araştırmada kullanılan veriler 2017 yılında paylaşıma açılan ve OECD'nin resmi internet sayfası olan http://www.oecd.org/pisa/data/2015database/ adresinden elde edilmiştir. SPSS veri dosyası formatında yer alan öğrenci anketinden ülke kodu TR olan 5895 öğrenciye ilişkin veriler analiz kapsamında veri kaynağı olarak kullanılmıştır.

Verilerin Analizi

Araştırmanın ilk aşamasında amaç öğrencilerin PISA verilerini kullanarak fen okuryazarlığı performanslarını tahmin edecek bir model oluşturmaktır. Öğrencilerin PISA sınavında gösterdikleri performans "Başarılı" ve "Başarısız" şeklinde kodlandığı için bu bir sınıflama problemidir ve veri madenciliği sınıflama yöntemleri kullanılarak analiz edilmiştir (Romero & Ventura, 2013). Bu sebeple literatürde sınıflama ve tahmin amacıyla sıklıkla kullanılan ve WEKA programında karar ağacı oluşturmaya izin veren Hoeffding Tree, J.48, Lojistik Model, REPTree ve Random Tree olmak üzere toplam 5 farklı öğrenme yöntemi kullanılarak analizler gerçekleştirilmiştir. Çalışmanın ikinci aşamasında farklı yöntemler yardımıyla elde edilen karar ağaçlarında anlamlı etkiye sahip olan değişkenler belirlenerek her bir karar ağacı için doğru sınıflama oranları üzerinden başarı oranları karşılaştırılacaktır. Bunun yanında modelleri karşılaştırmada kullanılan diğer ölçütler Kappa istatistiği, Mutlak hatanın ortalaması, Hataların ortalama karekökü, Göreceli mutlak hata ve Göreceli hataların karekökü olarak belirlenmiştir. Bu değerlerin karşılaştırılmasında standart bir ölçüt olmaması sebebiyle hataların olabildiğince düşük ve Kappa istatistiği ile doğru sınıflama oranının olabildiğince yüksek olması gerekmektedir (Almuniri & Said, 2017; Doreswamy, 2012; Hossin & Sulaiman, 2015; Kiranmai & Damodaram, 2014; Sokolova & Lapalme, 2009; Tiwari, Jha, & Yadav, 2012; Vihinen, 2012). Kappa istatistiği ya da Kappa değeri beklenen ve gözlenen değerlerin karşılaştırıldığı bir sayısal değer olup tesadüfi şans faktörünü de hesaba kattığından daha az yanıltıcı bir değer olarak kabul edilmektedir (Witten ve Frank, 2005). Bu süreçte gözlenen doğruluk oranı tüm matris üzerinden doğru olarak sınıflanmış örneklerin sayısının toplam örnek sayısına bölümü sonucunda elde edilmektedir. Beklenen doğruluk ise verilen matrise dayalı olarak herhangi bir tesadüfi sınıflandırıcının başarılı olması anlamına gelmektedir. Son olarak beklenen ve gözlenen doğruluk oranlarının belirlenmesinin ardından Kappa istatistiği aşağıda verilen eşitlik yardımıyla hesaplanmaktadır (Carletta, 1993).

$$\text{Kappa } (\kappa) = \frac{\text{Gözlenen doğruluk değeri} - \text{Beklenen doğruluk değeri}}{1 - \text{Beklenen doğruluk değeri}}$$

Böylece Kappa istatistiği makine öğrenme yöntemi tarafından tesadüfi bir sınıflandırıcının beklenen doğruluk değerini kontrol ederek veri setinde yer alan örneklerin gerçeğe ne kadar yakın bir şekilde sınıflandırıldığının ölçüsünü vermektedir. Kappa istatistiğinin yorumlanması konusunda kesin standart bir yorum olmasa da genel olarak .00-.20 arası düşük, .21-.40 arası kayda değer, .41-.60 arası orta düzeyde; .61-.80 arası önemli ve .81-1.00 arası mükemmel olarak tanımlanmaktadır (Landis & Koch, 1977).

Verilerin Analizinde Kullanılan Yazılım ve Algoritmalar

Verilerin analizinde Java tabanlı WEKA 3.8 (Waikato Environment for Knowledge Analysis) paket programından yararlanılmıştır. WEKA programı tarımsal verinin işlenmesi amacıyla Yeni Zelanda'daki Waikato Üniversitesi tarafından geliştirilmiştir (Kuyucu, 2012). Verilerin analizinde WEKA programının seçilmesinin temel sebepleri programın farklı alanlarda yaygın olarak kullanılmaya başlaması ve sistemin açık kaynak kodlu olmasıdır.

En popüler karar ağaçlarından biri olan ID3 Austurya Sidney üniversitesinden J.Ross Quinlen tarafından temeli atılan ve bilgi kazanım ölçütüne (information gain criterion) dayanan kararlı bir yöntemdir. Daha sonrasında bağımlı değişkenin sayısal olması, kayıp veriler, gürültülü veriler (noisy data) ve ağaçlardan kural oluşturma gibi birçok yenilik ve iyileştirmeye sahip C4.5 ağaç yöntemi ID3 esas alınarak geliştirilmiştir.

Lojistik Model Ağaçları (LMT) hedef veya sonuç değişkeni olarak tanımlanan özelliğin iki kategorili veya çok kategorili, sayısal veya sınıflama düzeyinde özellikler ve kayıp veriler ile çalışabilmektedir. Bir düğümde lojistik regresyon fonksiyonunu uyguladığınızda, analiz için kaç iterasyon yapılacağını belirlemek için çapraz geçerleme yöntemi kullanılır ve böylece her düğümde farklı bir iterasyon yerine ağacın tamamı boyunca aynı sayı kullanılmaktadır. Bu işlem çalışma süresini önemli ölçüde artırırken elde edilen sonuçların doğruluğu üzerinde çok az bir etkiye sahiptir. Alternatif olarak ağaç boyunca kullanılacak iterasyon sayısını kendiniz ayarlayabilirsiniz. Normalde çapraz geçerlik sayısının azaltılması yanlış sınıflandırma hatasıdır. Fakat bunun yerine olasılıkların ortalama hatalarının karekökü de seçilebilir. Ağaçtaki bölünme ölçütü C4.5'in bilgi miktarına dayalı olarak ya da lojit artık değerlerine göre belirlenmektedir. Burada artık değerlerin saflığını artırmak için uğraşılmaktadır.

Regresyon ağacı (REPTree) bilgi kazanım ve varyans azaltma yöntemiyle bir karar veya regresyon ağacı oluşturmaktadır. Varyans azaltması indirgenmiş hata budaması yöntemiyle gerçekleşmektedir. Bu yöntem C4.5 gibi eksik verileri parçalara ayırmak suretiyle üstesinden gelebilmektedir. Her bir yaprakta yer alacak en düşük örnek sayısı kullanıcılar tarafından belirlenebilir. Maksimum ağaç derinliği dallanma için eğitim setinin minimum oranı ve dallanma için düzey sayısı da seç düğmesinin yanındaki komut satırı tıklanarak açılan pencerede değişiklik yaparak istenildiği şekilde düzenlenebilmektedir.

Rastgele ağaç (RandomTree) yönteminde her bir düğümde daha önce belirlenen sayıda rastgele özellik seçilerek bir test yapılırken ağacın dallarında budama yapılmamaktadır. Karar ağaçlarında her bir düğümde en iyi çalışan yani en çok bilgi veren özellik seçilirken rastgele ağaç yönteminde bu seçim tesadüfi olarak gerçekleşmektedir (Witten et al., 2016).

Bulgular

PISA Fen okuryazarlığını yordamak amacıyla kullanılacak değişken sayısının ilgili literatür taraması sonucunda 29 olarak belirlenmesi sebebiyle ilk olarak farklı algoritmalar yardımıyla en iyi özellikler belirlenmeye çalışılmıştır. Çalışma kapsamında öğrencilerin fen okuryazarlığı üzerinde etkili olduğu belirlenen ve çalışma kapsamında kullanılan değişkenlerin isimleri, kodları, en düşük ve en yüksek değerleri ile her bir değişkenin 10 katlı çapraz geçerleme yönteminde başarılı olduğu katman sayılarına ilişkin yüzdelik değerler Tablo 1'de gösterilmiştir. Her ne kadar Yapay Sinir Ağları (YSA) yöntemleri regresyon analizi yöntemine kıyasla modele eklenen değişken sayısının fazla olması durumunda daha iyi tahmin yapacağı belirtilse de (Lykourentzou et al., 2009); VM gibi tahmine dayalı yöntemlerde modele daha çok sayıda değişken eklemenin performans kestiriminin doğruluğunda bir artışa sebep olmayacağı belirtilmiştir (Huang & Fang, 2013). Bunun yanında Kohavi (1995) tarafından yapılan deneysel çalışmada farklı algoritmalar altında yapılan kestirimlerin katman sayısının 10 ve 20 olması durumunda oldukça iyi olduğu ve neredeyse tamamen yansız olduğu belirtildiği için çalışmada 10 katlı çapraz geçerleme yöntemi kullanılmıştır. Bu işlemlerin sonucunda 12 bağımsız değişken tarafından PISA fen okuryazarlığını yordamak amacıyla kullanılan veri madenciliği yöntemlerinin doğru sınıflama sayıları, doğru sınıflama oranları, güvenirlik değerleri ve elde edilen karar ağaçlarının karşılaştırılması aşamasına geçilmiştir. Bu aşamada WEKA programında PISA fen okuryazarlığı kategorik değişkene dönüştürülmüştür. Öğrencilerin

PISA okuryazarlık puanları bakımından Türkiye ortalaması olan 425.00 puanının altında olanlar başarısız (0), bu puanın üzerinde olanlar başarılı (1) olarak kodlanmıştır.

Tablo 1.
Veri Madenciliğinde Kullanılacak Değişkenlere İlişkin Özet Bilgiler.

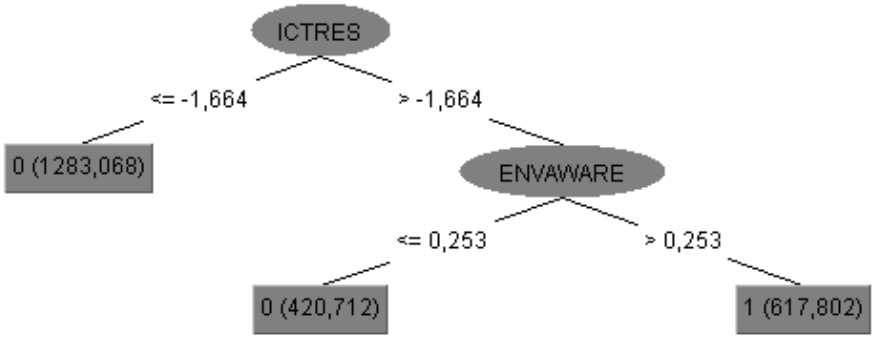
Sıra	Değişkenin Adı	Kodu	min	max	Başarı oranı
1	Sorgulamaya dayalı fen eğitimi	IBTEACH	-3.34	3.18	%100.00
2	Çevreye duyarlık	ENVAWARE	-3.38	3.29	%100.00
3	Bilimsel inançlar	EPIST	-2.79	2.16	%100.00
4	Öğrenciden beklenen mesleki statü	BSMJ	10.00	89.00	%100.00
5	Okul dışı ders çalışma süresi (hafta)	OUTHOURS	.00	70.00	%100.00
6	Fen öğrenme süresi (hafta)	SMINS	.00	800.00	%100.00
7	Toplam öğrenme süresi (hafta)	TMINS	100.00	3000.00	%100.00
8	Test kaygısı	ANXTEST	-2.51	2.55	%60.00
9	Öğretmenin adil olması	unfairteacher	1.00	24.00	%40.00
10	Evdeki eğitim kaynakları	HEDRES	-4.37	1.18	%80.00
11	BİT kaynakları	ICTRES	-3.27	3.50	%100.00
12	Sosyo ekonomik durum indeksi	ESCS	-5.13	3.12	%100.00

Farklı Öğrenme Yöntemlerine İlişkin Bulgular

PISA öğrenci anketiyle ölçülen değişkenlerden yararlanarak başarıyı tahmin etmede kullanılan, Hoeffding Tree, J.48, Lojistik Model, RepTree ve Random Tree yöntemleri yardımıyla edilen ölçme sonuçlarının 10 katlı çapraz geçerleme yöntemi yardımıyla elde edilen sonuçlar sırasıyla rapor edilmiştir.

Hoeffding Tree Yöntemine İlişkin Bulgular

Hoeffding Tree yöntemiyle toplam 5865 öğrencinin 4085 tanesi doğru sınıflanarak bu yöntemle elde edilen doğru sınıflama oranı %69.65 olarak belirlenmiştir. Sınıflamaya ilişkin kappa istatistiğinin .39 ve ortalama mutlak hatanın .36 olduğu tespit edilmiştir. Hoefding tree yöntemi için program tarafından oluşturulan ağaç yapısı Şekil 2'de gösterilmiştir.



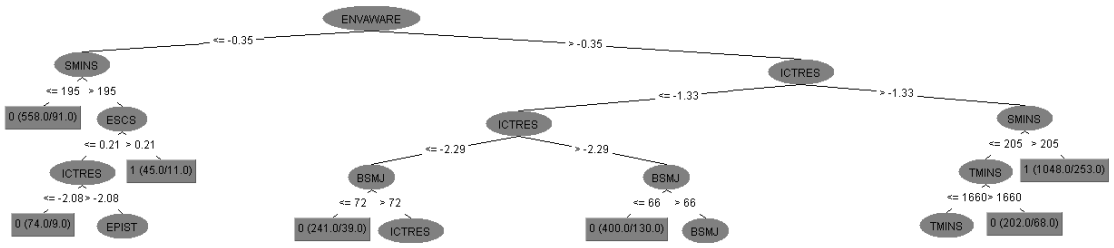
Şekil 2. Hoeffding tree yöntemiyle elde edilen ağaç yapısı.

PISA fen okuryazarlığı bakımından öğrencileri sınıflamada ilk olarak Bilgi ve İletişim Teknolojileri kaynakları ve sonrasında çevreye duyarlık değişkenlerinin etkili olduğu belirlenmiştir. Elde edilen ağaç yapısı program tarafından üretilmiş olup daha alt düzeylerde dallanma gerçekleşmemiştir. Başka bir ifadeyle elde edilen karar ağacı iki düzeyli olarak belirlenmiştir. Karar ağacında Bilgi ve İletişim Teknolojisi kaynakları (ICTRES) bakımından kesme puanının -1.66 olduğu ve bu değerin altında kalanların başarısız olarak sınıflanırken bu değerin üzerinde olanların çevreye duyarlık puanlarına bakıldığı ve çevreye duyarlık için kesme puanının .25 olduğu belirlenmiştir. Sınıflamada kullanılan saflığın ölçüsü bilgi

(information) olarak adlandırılır ve BİT olarak belirlenen birimlerle ölçülür. Entropi olarak adlandırılan bilgi miktarları esas alındığında çalışma kapsamında ele alınan 12 değişkenden sadece 2 tanesinin öğrencileri başarılı ve başarısız olarak sınıflamada yeterli olduğu sonucuna ulaşılmaktadır.

J.48 Yöntemine İlişkin Bulgular

J.48 yöntemiyle toplam 5865 öğrencinin 3976 tanesi doğru sınıflanarak bu yöntemle elde edilen doğru sınıflama oranı %67.79 olarak belirlenmiştir. Sınıflamaya ilişkin Kappa istatistiğinin .35 ve ortalama mutlak hatanın .38 olduğu tespit edilmiştir. J.48 yöntemi için program tarafından oluşturulan ağaç yapısının bir kısmı Şekil 3'te gösterilmiştir.



Şekil 3. J.48 yöntemiyle elde edilen ağaç yapısı.

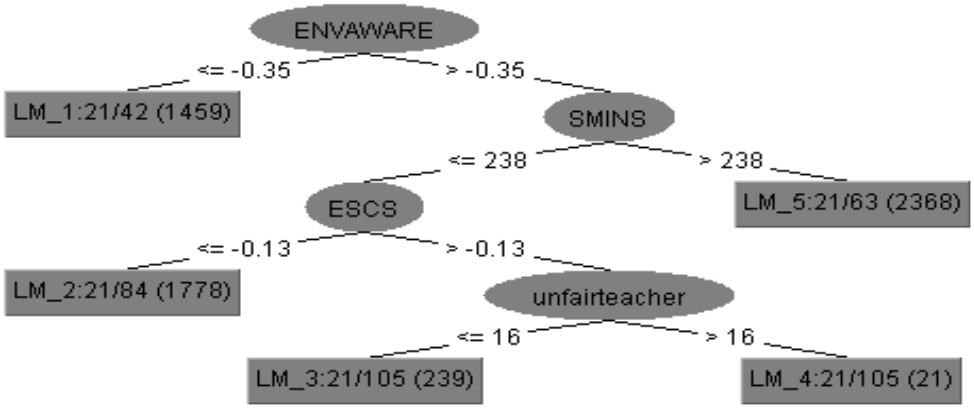
J.48 yöntemiyle elde edilen karar ağacı toplam 15 düzeyli olup yorumlamada kolaylık olması amacıyla ilk dört düzeyi gösterilmiştir. PISA fen okuryazarlığı bakımından öğrencileri sınıflamada ilk olarak çevreye duyarlık ve sonrasında fen öğrenme süresi ile BİT kaynakları değişkenlerinin etkili olduğu belirlenmiştir. Bir alt düzeydeki sınıflamada ise sosyo ekonomik durum indeksi ile yine fen öğrenme süresi ve BİT kaynakları değişkenlerinin etkili olduğu görülmektedir. Çevreye duyarlık bakımından kesme puan program tarafından ham puanlar üzerinden hesaplanmış ve -.35 olduğu belirlenmiştir. Bu değerin altında kalanların fen öğrenme sürelerine bakıldığı ve fen öğrenme süresi için kesme puanının 195.00 olduğu belirlenmiştir. Haftalık fen öğrenme süresi 195.00 dakikanın altında olanlar başarısız olarak sınıflanırken fen öğrenme süresi 238.00 dakikadan fazla olanları sınıflamada sosyo ekonomik durum indeksi değişkeninin etkili olduğu görülmektedir. Karar ağacının aynı dalı üzerinde bir alt düzeyde sosyo ekonomik durum indeksi değişkeni için kesme puanı .21 olarak belirlenmiş ve bu değerin üzerinde olanların başarılı olarak sınıflanırken bu değerin altında olanlarda BİT kaynakları değişkenine bakıldığı belirlenmiştir.

Lojistik Model Yöntemine İlişkin Bulgular

Lojistik Model (LM) yöntemiyle toplam 5865 öğrencinin 4125 tanesi doğru sınıflanarak bu yöntemle elde edilen doğru sınıflama oranı %70.33 olarak belirlenmiştir. Sınıflamaya ilişkin Kappa istatistiğinin .40 ve ortalama mutlak hatanın .38 olduğu tespit edilmiştir. Lojistik Model yöntemi için program tarafından oluşturulan ağaç yapısının bir kısmı Şekil 4'te gösterilmiştir.

Elde edilen ağaç yapısı program tarafından üretilmiş olup daha alt düzeylerde dallanma gerçekleşmemiştir. Başka bir ifadeyle elde edilen karar ağacı üç düzeyli olarak belirlenmiştir. PISA fen okuryazarlığı bakımından öğrencileri sınıflamada ilk olarak çevreye duyarlık ve sonrasında haftalık fen öğrenme süresi özelliğinin etkili olduğu belirlenmiştir. Bir alt düzeydeki sınıflamada ise sosyo ekonomik durum indeksi ve son alt basamakta öğretmenin adil olması değişkeninin etkili olduğu görülmektedir. Çevreye duyarlık bakımından kesme puanının -.35 olduğu ve bu değerin altında kalanların 1 numaralı lojistik modele göre sınıflanırken bu değerin üzerinden olanların haftalık fen öğrenme sürelerine bakıldığı ve fen öğrenme süresi için kesme puanının 238.00 olduğu belirlenmiştir. Fen öğrenme süresi 238.00 saatin altında olan öğrencileri sınıflamada Sosyo ekonomik durum indeksi etkili olurken fen öğrenme süresi 238 saatin üzerinde olanlar doğrudan LM_5 isimli modele göre sınıflanmaktadırlar. Elde edilen bu sonuç J.48 yöntemi ile benzerlik göstermektedir. Şekil 4 incelendiğinde karar ağacının yapraklarında LM_1, LM_2, LM_3, LM_4 ve LM_5 şeklinde beş farklı lojistik modele dayalı denklem oluşturulduğu

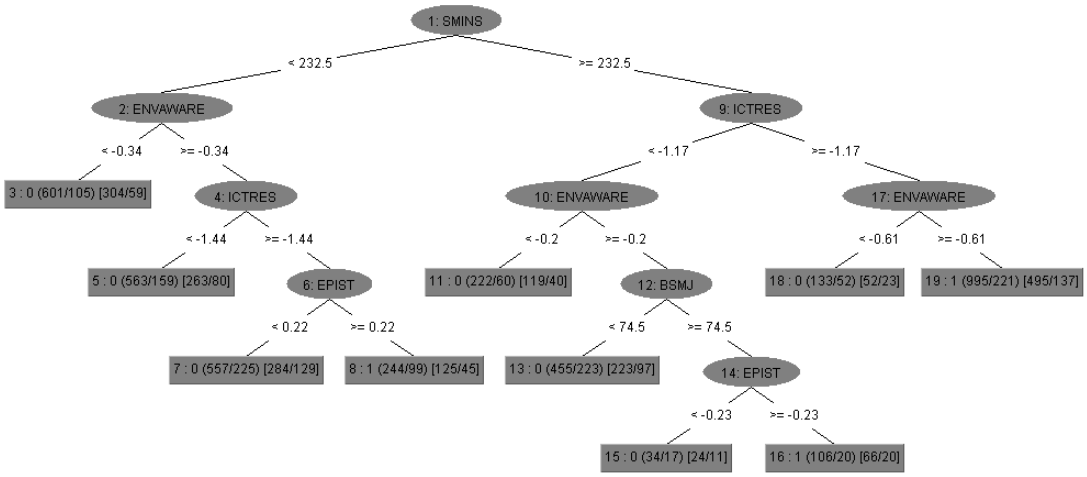
görülmektedir. Geleneksel istatistiksel analizlerde bağımlı değişkenin bağımsız değişkenler tarafından yordanmasına ilişkin tek bir regresyon denklemi oluşturulurken Lojistik modelde beş farklı regresyon denklemi elde edilmektedir. Bu nedenle lojistik denklemin başarı oranı diğer yöntemlerden daha yüksek olmaktadır. Bunun yanında her bir lojistik denklemde sabit sayıların ve değişkenlerin önünde yer alan katsayıların farklılık gösterdiği belirlenmiştir.



Şekil 4. Lojistik model yöntemiyle elde edilen ağaç yapısı.

REPTree Yöntemine İlişkin Bulgular

REPTree toplam 5865 öğrencinin 3946 tanesi doğru sınıflanarak bu yöntemle elde edilen doğru sınıflama oranı %67.28 olarak belirlenmiştir. Sınıflamaya ilişkin Kappa istatistiğinin .34 ve ortalama mutlak hatanın .39 olduğu tespit edilmiştir. REPTree yöntemi için program tarafından oluşturulan ağaç yapısının bir kısmı Şekil 5'te gösterilmiştir.

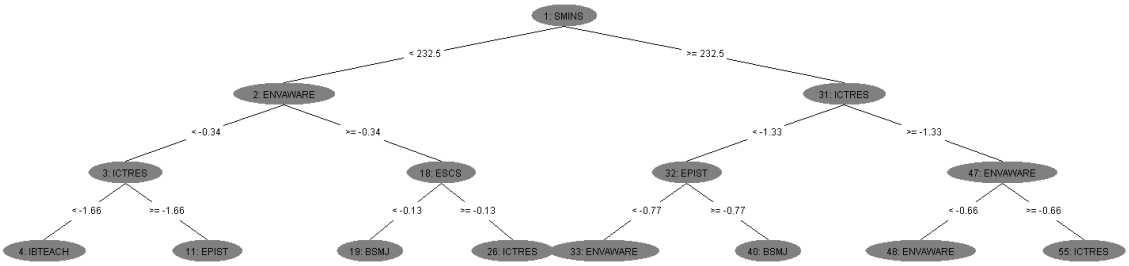


Şekil 5. *REPTree yöntemiyle elde edilen ağaç yapısı.*

REPTree yöntemiyle elde edilen karar ağacı toplam 13 düzeyli olup yorumlamada kolaylık olması amacıyla ilk üç düzeyi gösterilmiştir. PISA fen okuryazarlığı bakımından öğrencileri sınıflamada ilk olarak haftalık fen öğrenme süresi ve sonrasında çevreye duyarlık ile BİT kaynakları özelliklerinin etkili olduğu belirlenmiştir. Bir alt düzeydeki sınıflamada ise çevreye duyarlık değişkeninin etkili olduğu görülmektedir. Ağacın ilk sınıflamasında haftalık fen öğrenme süresi bakımından kesme puanının 232.50 olduğu ve bu değerin altında kalanların çevreye duyarlık puanlarına bakılırken bu değerin üzerinde olanların BİT kaynakları puanına bakılmaktadır. Benzer şekilde belirtilen her iki dal üzerinde BİT kaynakları için belirlenen kesme puanları sırasıyla -1.67 ve -1.44 olarak belirlenmiştir.

Random Tree Yöntemine İlişkin Bulgular

Random Tree yöntemiyle toplam 5865 öğrencinin 3690 tanesi doğru sınıflanarak bu yöntemle elde edilen doğru sınıflama oranı %62.92 olarak belirlenmiştir. Sınıflamaya ilişkin Kappa istatistiğinin .26 ve ortalama mutlak hatanın .37 olduğu tespit edilmiştir. Random Tree yöntemi için program tarafından oluşturulan ağaç yapısının bir kısmı Şekil 6'da gösterilmiştir.



Şekil 6. Random tree yöntemiyle elde edilen ağaç yapısı.

Random tree yöntemiyle elde edilen karar ağacı toplam 24 düzeyli olup yorumlamada kolaylık olması amacıyla ilk üç düzeyi gösterilmiştir. PISA fen okuryazarlığı bakımından öğrencileri sınıflamada ilk olarak haftalık fen öğrenme süresi ve sonrasında çevreye duyarlık ile BİT kaynakları değişkenlerinin etkili olduğu belirlenmiştir. Bir alt düzeydeki sınıflamada ise BİT kaynakları, bilimsel inançlar, çevreye duyarlık ve sosyo ekonomik durum indeksi değişkenlerinin etkili olduğu görülmektedir. Ağacın ilk sınıflamasında haftalık fen öğrenme süresi bakımından kesme puanının 232.50 olduğu ve bu değerin altında kalanların çevreye duyarlık puanlarına bakılırken bu değerin üzerinde olanların BİT kaynakları puanına bakılmaktadır. Çevreye duyarlık için kesme puanı -.34 olarak belirlenmiş ve bu değerin altında olanları sınıflamada BİT kaynakları etkili olurken bu değerin üzerinde olanları sınıflamada sosyo ekonomik durum indeksi değişkeni etkili olmaktadır.

Farklı Yöntemlerin Karşılaştırılması

Farklı yöntemler yardımıyla elde edilen güvenirlik değerlerinin karşılaştırılması amacıyla ağırlıklandırılmış genel ortalama değerleri karşılaştırılmıştır. Her bir yöntem için doğru sınıflanan örnek sayıları, doğru sınıflama oranları, Kappa istatistiği, Kappa düzeyi ve mutlak hatanın ortalaması değerleri Tablo 2'de gösterilmiştir.

Tablo 2.
Farklı Yöntemler Kullanılarak Elde Edilen Güvenirlik Ölçütleri.

Algoritma	Sınıflamada kullanılan düzey sayısı	Doğru sınıflama sayısı	Doğru sınıflama oranı	Kappa istatistiği	Kappa düzeyi	Ortalama mutlak hata
1.Hoefding tree	2	4085	69.65	.39	Kayda değer	.36
2. J.48	15	3976	67.79	.35	Kayda değer	.38
3. Lojistik model	3	4125	70.33	.40	Orta düzeyde	.38
4. REPTree	13	3946	67.28	.34	Kayda değer	.39
5. Random tree	24	3690	62.92	.26	Kayda değer	.37

Tablo 2 incelendiğinde doğru sınıflama sayısı, doğru sınıflama oranı, Kappa istatistiği ve Kappa düzeyi değerleri bakımından en iyi sonuçların Lojistik model yöntemiyle elde edilirken ortalama mutlak hata bakımından en iyi sonuçların Hoefding Tree yöntemiyle elde edildiği belirlenmiştir. Eğer Kappa istatistiği dikkate alınarak bir sıralama yapmak istenirse en iyi sonuçların sırasıyla Lojistik model, Hoefding tree, J.48, RepTree ve Random Tree algoritmaları şeklinde olacağı görülmektedir. Buna göre doğru sınıflama veya kappa istatistiğini esas almanın sıralamada bir değişikliğe neden olmayacağı görülmektedir.

Tartışma, Sonuç ve Öneriler

Araştırmada fen okuryazarlığını yordamak amacıyla kullanılan değişken sayısının PISA öğrenci anketi esas alındığında çok fazla olması sebebiyle öncelikle çalışma kapsamında kullanılacak değişkenler belirlenmiştir. Değişken sayısını azaltmak amacıyla öncelikle VM yöntemlerinden Best First Forward, Best First Backward ve Greedy Stepwise yardımıyla 10 katlı çapraz geçerleme yöntemi sonucunda en az %40.00 ve üzeri başarılı olan toplam 12 değişken yardımıyla PISA fen okuryazarlığı tahmin edilmeye çalışılmıştır. Çalışmada farklı yöntemler tarafından elde edilen karar ağaçlarına ait düzey sayısının doğru sınıflama oranlarına doğrudan bir etkiye sahip olmadığı belirlenmiştir.

Bakker (2016) aynı veri seti üzerinden farklı yöntemler yardımıyla elde edilen karar ağaçlarının farklılık gösterebileceğini belirtmektedir. Nitekim çalışma kapsamında fen okuryazarlığı en iyi yordayan değişken Reptree ve random tree yöntemlerinde fen öğrenme süresi iken J.48 ve lojistik modelde çevreye duyarlık, Hoefding tree yönteminde ise Bilgi ve İletişim Teknolojileri kaynakları olarak belirlenmiştir. Her ne kadar Reptree ve Random Tree yöntemlerinde en çok bilgi veren özellik fen öğrenme süresi olarak belirlense de bu iki yöntem tarafından elde edilen karar ağacının bir alt düzeydeki dallanmasında Reptree yönteminde çevreye duyarlık ve BİT kaynakları etkili iken Random Tree yönteminde çevreye duyarlık ve evdeki eğitim kaynaklarının etkili olduğu belirlenmiştir. Benzer şekilde diğer yöntemler için elde edilen karar ağaçlarında da farklılık olduğu göze çarpmaktadır. Elde edilen bu sonuç, Mohan (2013) tarafından belirtilen, bir yöntemin kendi içinde parametrelerini değiştirdiğinizde duyarlık, hassaslık ve geri getirme gibi ölçütlerinde bir farklılık oluşmazken farklı yöntemler kullanıldığında elde edilen sonuçlarda farklılık olacağı bulgusuyla benzerlik göstermektedir.

Çalışmada farklı karar ağacı yöntemlerinin doğru sınıflama oranları karşılaştırıldığında sırasıyla Lojistik model, Hoefding tree, J.48, RepTree ve Random Tree yöntemlerinin en başarılı yöntemler olduğu belirlenmiştir. Elde edilen bu sonuca göre kullanılan karar ağacına göre doğru sınıflama oranları da farklılık göstermektedir. Bu sonuç C4.5, CART ve Random forest yöntemlerinin başarı oranları arasında önemli farklar olduğunu gösteren Cinaroglu (2016) tarafından yapılan çalışma ile benzerlik göstermektedir. Benzer bir çalışmada Dietterich (2000) karar ağacı oluşturmada kullanılan, yerine koyarak örnekleme (bagging), ardışık topluluklarla öğrenme (boosting) ve rastsal öğrenme (randomization) yöntemlerinin elde edilen sonuçlarda farklılık yaratacağını belirlemiştir. Bu yönüyle çalışmanın bulguları benzerlik göstermektedir.

Çalışmada güvenirlik ölçütlerinden doğru sınıflanan örnek sayısı, kappa istatistiği ve ortalama mutlak hata değerlerine göre en iyi tahmin yönteminin Lojistik model olduğu belirlenmiştir. Elde edilen bu sonuç Liaw ve Wiener (2002) ile Svetnik, Liaw, Tong ve Wang (2004) tarafından yapılan çalışmalarla benzerlik göstermektedir. Lojistik model yardımıyla elde edilen karar ağacında tek bir regresyon denklemi yerine birden çok regresyon denkleminin kullanılması ve sonrasında onların tahminlerinden elde edilen sonuçlar ile yeni veriyi sınıflandıran öğrenme algoritmaları oluşturulduğu için daha az hatalı ve daha güvenilir sonuçlar elde edildiği düşünülmektedir (Strobl et al., 2009). Daha sonra sırasıyla Hoeffding Tree, J4.8, RepTree ve son olarak Random Tree yöntemlerinin en az hataya sahip ve en güvenilir yöntemler olduğu belirlenmiştir. Alanyazında en çok kullanılan yöntem olarak görülen J4.8 algoritması orta düzeyde bir başarı göstermiştir. Bunun yerine lojistik model ve Hoeffding tree yöntemlerinin hem hataya dayalı güvenirlik değerleri hem de geçerlik ölçütleri bakımından daha başarılı yöntemler olduğu belirlenmiştir.

Öneriler

Çalışmada farklı yöntemlerle elde edilen karar ağaçlarında çıktı değişkeni üzerinde etkisi olan girdi değişkenlerin farklılık gösterdiği belirlenmiştir. Elde edilen bu bulgulara dayalı olarak akademik başarının yordanmasına ilişkin çalışmalarda veri madenciliğinde tek bir yöntemle çalışmak yerine en az 3 farklı öğrenme yöntemi kullanarak elde edilen sonuçların geçerliğine ilişkin delil sunulmalıdır.

Veri madenciliğinde özellikle WEKA programında yer alan öğrenme yöntemlerinden lojistik model (LMT) ve Hoefding tree yöntemlerinin farklı koşullar altında genellikle daha güvenilir ve geçerli sonuçları elde ettiği belirlendiğinden başarının yordanması veya sınıflama amacıyla yapılacak çalışmalarda bu öğrenme yöntemlerinin kullanılması önerilmektedir. Özellikle lojistik modelde başarılı ve başarısız olarak sınıflanan öğrencilere ilişkin elde edilecek karar ağacının yapraklarında yer alan lojistik denklemler yardımıyla değişkenler arasındaki ilişki tek bir regresyon denklemi ile modellemek yerine farklı özellik değerlerine göre çok sayıda regresyon denklemi kullanarak doğru sınıflama oranı artırılabilir.

Çalışmada PISA verileri kullanarak analiz yapıldığından ileride ÖSYM tarafından yapılan geniş ölçekli sınavlardan elde edilen sonuçların korelasyon ve regresyona dayalı standart istatistiksel teknikler yerine daha zengin kanıtlar sunmak amacıyla veri madenciliğine dayalı yöntemlerle analiz edilmesi önerilebilir

References

Almuniri, I., & Said, A. M. (2017). School's performance evaluation based on data mining. *International Journal of Engineering and Information Systems*, *1* (9), 56–62.

Aydın, S. (2007). *Veri madenciliği ve Anadolu Üniversitesi uzaktan eğitim sisteminde bir uygulama*. Unpublished doctorate dissertation, Anadolu Üniversitesi, Eskişehir.

Bakker, R. (2016). *A comparison of decision trees for ingredient classification*. Bachelor thesis, University of Amsterdam, Amsterdam.

Barros, R. C., Carvalho, A. C. P. L. F. de, & Freitas, A. A. (2015). *Automatic design of decision-tree ınduction algorithms*. Heidelberg, NY: SpringerBriefs in Computer Science.

Büyüköztürk, Ş., Kılıç Çakmak, E., Akgün, Ö., Karadeniz, Ş., & Demirel, F. (2016). *Eğitimde bilimsel araştırma yöntemleri*. Ankara: Pegem Akademi Yayıncılık.

Carletta, J. (1993). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, *22* (2), 249–254.

Cinaroglu, S. (2016). Comparison of performance of decision tree algorithms and random forest: An application on OECD countries health expenditures. *International Journal of Computer Applications*, *138* (1), 37–41.

Dietterich, T. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning*, *40* (2), 139–157.

Doreswamy, H. K. (2012). Performance evaluation of predictive classifiers for knowledge discovery from engineering materials data sets. *CIIT International Journal of Artificial Intelligent Systems and Machine Learning*, *3* (3), 162–168.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, *39* (11), 27–34. https://doi.org/10.1145/240455.240464

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2012). *How to design and evaluate research in education*. New York: McGraw-Hill.

Han, J., & Kamber, M. (2006). *Data mining concepts and techniques*. San Francisco, CA: Morgan Kaufmann Publishers.

Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process*, *5*(2), 01–11. https://doi.org/10.5121/ijdkp.2015.5201

Hssina, B., Abdelkarim, M., Ezzikouri, H., & Erritali, M. (2014). A comparative study of decision tree ID3 and C4.5. *International Journal of Advanced Computer Science and Applications*, *4* (2), 13–19. https://doi.org/10.14569/specialissue.2014.040203

Huang, S., & Fang, N. (2013). Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. *Computers and Education*, *61* (1), 133–145. https://doi.org/10.1016/j.compedu.2012.08.015

Imielinski, T., & Mannila, H. (1996). A database perspective on knowledge discovery. *Communications of the ACM*, *39* (11), 373–408.

Kiranmai, B., & Damodaram, A. (2014). *A review on evaluation measures for data mining tasks*. *International Journal Of Engineering And Computer Science, 3* (7), 7217–7220.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, 2, 1137–1145. https://doi.org/10.1067/mod.2000.109031

Kusiak, A. (2001). Data analysis: Models and algorithms. *Proceedings of the SPIE Conference on Intelligent Systems and Advanced Manufacturing*, In P.E. Orban and G.K. Knopf (Eds), *SPIE* (pp. 1-9), Boston: MA.

Kuyucu, Y. E. (2012). *Lojistik regresyon analizi (LRA), yapay sinir ağları (YSA) ve sınıflandırma ve regresyon ağaçları (C&RT) yöntemlerinin karşılaştırılması ve tıp alanında bir uygulama*. Unpublished master's thesis, Gaziosmanpaşa Üniversitesi, Sağlık Bilimleri Enstitüsü, Tokat.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33* (1), 159–174. https://doi.org/10.2307/2529310

Larose, D. T. (2005). *Discovering knowledge in data: An introduction to data mining*. New Jersey: John Wiley & Sons.

Liaw, A., & Wiener, M. (2002). Classification and regression by random forest. *R News*, *2* (3), 18–22.

Lv, S., Kim, H., Zheng, B., & Jin, H. (2018). A review of data mining with big data towards its applications in the electronics industry. *Applied Sciences*, *7* (582), 2–34. https://doi.org/10.3390/app8040582

Lykourentzou, I., Giannoukos, I., Mpardis, G., Nikolopoulos, V., & Loumos, V. (2009). Early and dynamic student achievement prediction in e-learning courses using neural networks. *Journal of the American Society for Information Science and Technology*, *60* (2), 372–380.

Maimon, O., & Rokach, L. (2005). *Data mining and knowledge discovery handbook*. Secaucus, NJ: Springer-Verlag Inc.

Mease, D., & Wyner, A. (2008). Evidence contrary to the statistical view of boosting: A rejoinder to responses. *Journal of Machine Learning Research*, *9*, 195–201.

MEB. (2016). *PISA 2015 Ulusal Raporu*. Millî Eğitim Bakanlığı, Ölçme, Değerlendirme ve Sınav Hizmetleri Genel Müdürlüğü, Ankara.

Mohan, V. (2013). Decision trees: A comparison of various algorithms for building decision trees. Retrieved July 23, 2019, from https://pdfs.semanticscholar.org/3399/c175beca3ab4843d67f91bb28f564099d0bb.pdf

Neelamegam, S., & Ramaraj, E. (2013). Classification algorithm in data mining: An Overview. *International Journal of P2P Network Trends and Technology (IJPTT)*, *3* (5), 1–5.

Podgorelec, V., Kokol, P., Stiglic, B., & Rozman, I. (2002). Decision trees: An overview and their use in medicine. *Journal of Medical Systems*, *26* (5), 445-463.

Rastogi, R., & Shim, K. (2000). PUBLIC: A decision tree classifier that integrates building and pruning. *Data Mining and Knowledge Discovery*, *4* (4), 315–344.

Romero, C., & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *3* (1), 12–27. https://doi.org/10.1002/widm.1075

Sieber, J. E. (2008). Data mining: Knowledge discovery for human research ethics. *Journal of Empirical Research on Human Research Ethics*, *3* (3), 1–2. https://doi.org/10.1525/jer.2008.3.3.1

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, *45* (4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, *14* (4), 323–348. https://doi.org/10.1037/a0016973.

Svetnik, V., Liaw, A., Tong, C., & Wang, T. (2004). Application of breiman's random forest to modeling structure-activity relationships of pharmaceutical molecules. In F. Roli, J. Kittler, & T. Windeatt, (Eds.). *Multiple Classifier Systems* (pp. 1-35), Berlin, Heidelberg: Springer.

Tan, P.N., Steinbach, M., & Kumar, V. (2005). *Introduction to data mining*. Boston, USA: Addison-Wesley Longman Publishing Co.

Thuraisingham, B. (2003). *Web data mining and applications in business ıntelligence and counter terrorism*. USA: CRC Press LLC, Boca Raton, FL.

Tiwari, M., Jha, M. B., & Yadav, O. (2012). Performance analysis of data mining algorithms in weka. *IOSR Journal of Computer Engineering (IOSRJCE)*, *6* (3), 32–41.

Vaus, D. de. (2001). *Research design in social research*. London: Sage Publications.

Vihinen, M. (2012). How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. *BMC Genomics*, *13* (4), 1-10. https://doi.org/10.1186/1471-2164-13-S4-S2

Weiss, S. M., & Kulikowski, C. A. (1991). *Computer systems that learn: Classification and prediction methods from statistics, neural nets, machine learning, and expert systems*. San Mateo, CA: Morgan Kaufmann.

Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques*. San Francisco, CA: Morgan Kaufmann Publishers.

Witten, I. H., Frank, E., & Hall, M. (2016). *Data mining: Practical machine learning tools and techniques*. USA: Morgan Kaufmann Publications.

Wu, X., Kumar, V., Ross, Q. J., Ghosh, J., Yang, Q., Motoda, H., … Steinberg, D. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems, 14* (1), 1-37. https://doi.org/10.1007/s10115-007-0114-2