Orclever Proceedings of Research and Development
https://doi.org/10.56038/oprd.v3i1.376
OR CLEVER Science & Research Group

Conference Article

Data Mining, Weka Decision Trees

Zekeriya DURAN[1*], İsmail AKARGÖL[2*], Tuğba DOĞAN[3*]

[1] Sivas Cumhuriyet University, Orcid ID: https://orcid.org/0000-0002-9327-8567, zduran@cumhuriyet.edu.tr,
[2] Sivas Cumhuriyet University, Orcid ID: https://orcid.org/0000-0002-0721-7064, iakargol@cumhuriyet.edu.tr,
[3] Sivas Cumhuriyet University, Orcid ID: https://orcid.org/0000-0002-2628-4238, tcamuzcu@cumhuriyet.edu.tr,
* Correspondence: zduran@cumhuriyet.edu.tr; 0505 930 14 73

(First received September 15, 2023 and in final form December 24, 2023)

3rd International Conference on Design, Research and Development (RDCONF 2023)
December 13 - 15, 2023

Reference: Duran, Z., Akargöl, İ., Doğan, T. Data Mining, Weka Decision Trees**.** Orclever Proceedings of Research and Development,3(1), 401-416.

Abstract

Nowadays, computer technologies are increasing rapidly. Thanks to the development of computer technologies, large and complex raw data sets can be transformed into useful information with different analysis techniques. Different algorithms developed thanks to computer technologies can offer different solutions to scientists and users working in different branches of science, especially engineering sciences, mathematics, medicine, industry, financial/economic fields, marketing, education, multimedia and statistics. Thanks to these solutions, it is possible to easily achieve the desired goals and objectives. Thus, by correctly managing and analyzing existing data in large and complex raw data datasets, accurate predictions can be made to be used in similar problems in the future. Data sets are analyzed and evaluated using different methods. It is also possible that the classification of data during the analysis and evaluation stages of data sets significantly affects the decision-making process regarding the work to be done. Classification of data can be done by statistical method or data mining method. Decision trees, which can be used to classify numerical and alphanumeric data, generally provide a great advantage for decision makers in terms of easy interpretation and understandability compared to other classification techniques. For these

https://doi.org/10.56038/oprd.v3i1.376

reasons, in this study, decision trees, one of the most used classification techniques in data mining, are mentioned.

Keywords: Weka, decision trees, classification

1. Introduction

Data mining; Valid and useful information is obtained through different analysis techniques hidden in large data sets; [1,2,3,4]. It is defined as an interdisciplinary branch of research that combines techniques and algorithms from several disciplines, including computer science, mathematics, and statistics [5]. It is used for classification and prediction purposes in many different fields such as medicine, industry, financial/economic fields, engineering, marketing, education and multimedia (3,6,7). The steps followed in the data mining process generally consist of five stages: defining the problem, preparing the data, establishing and evaluating the model, using the model and monitoring the model [8,9]. After data sets are taken from many sources and combined and sent to data warehouses, the data in the warehouses are taken and pre-processed before being evaluated in a standard format. The processed data is transferred to a data mining algorithm that produces an output in the form of rules or another type of pattern and is interpreted according to the user's wishes with the help of algorithms (patterns) and converted into information. Its place in the knowledge discovery phase used in data mining is shown in Figure 1[10].

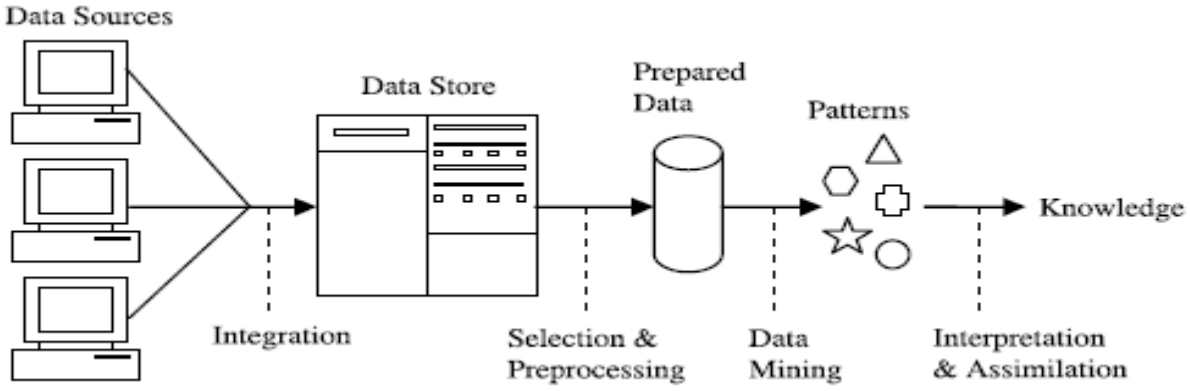

Figure 1. Data mining in the knowledge discovery phase [10]

https://doi.org/10.56038/oprd.v3i1.376

Nowadays, when discovering information from databases using advanced information technologies has become of great importance, data mining looks for hidden networks, patterns, correlations and interdependencies in huge databases that traditional information gathering (report creation, pie and bar chart creation, user query, decision support systems, etc.) methods may miss. Additionally, data mining consists of newly discovered algorithmic models that help solve user-defined problems quickly.; It also uses a mixed set of tools. Each model in these powerful toolkits; It possesses an inherent simplicity, making it easily comprehensible and user-friendly. These toolsets encompass a range of artificial intelligence techniques such as expert systems, fuzzy logic, decision trees, rule induction methods, genetic algorithms, genetic programming, neural networks, and clustering techniques. Furthermore, through the incorporation of data visualization, it serves as a valuable tool for facilitating the development, generation, and interpretation of knowledge derived from data-driven discoveries [11].

Data Mining mainly uses two different models: predictive and descriptive (2,3,12). For example; While the classification technique uses the regression predictive model, the association rule, clustering and sequential time models use the descriptive model. The classification technique is most commonly used in data mining, and it is a model used to predict data classes whose class is not yet known, using data with a defined class. This model consists of two stages. In the first stage of the model, a model is created for prediction purposes, and in the second stage, the model class desired by the user is estimated using data whose class is unknown. The most preferred classification techniques can be listed as Support Vector Machines, Artificial Neural Networks, , Genetic Algorithms, Logistic Regression, K-Nearest Neighbour and Decision Trees [4,12, 13, 14, 15, 16].

It is extremely important to make accurate predictions for use in similar problems in the future by correctly managing and analyzing existing data in large data sets. Data sets can be analyzed in various ways. By classifying the data, the decision-making process regarding the work to be done will be significantly affected. Classification of data can be done by statistical method or data mining method. The data classification process consists of two stages. The initial phase involves learning, during which a classification algorithm generates the classifier by examining (or acquiring knowledge from) a training set that includes databases along with their corresponding class labels. In this first step, the model or classifier derives the function $Y = f(X)$, which, depending on a variable X, can take the form of mathematical inequalities, decision trees, or classification rules. In the next stage,

https://doi.org/10.56038/oprd.v3i1.376

the classifier is predicted correctly. For this, the classifier given in the first step is tested with a group set other than the training groups, the "test set" [17, 18, 19].

2. Materials and Methods

2.1. WEKA Decision Trees

Upon reviewing the literature, it becomes evident that various data mining programs are employed. Among these; SAS Enterprise Miner, Rapid Miner, MATLAB, Salford Predictive Modeling Suite (SPM), STATISTICA, Orange, Knime, Python, Scikit-Learn and WEKA [20, 21, 22]. Software developed at the University of Waikato in New Zealand is formally known as the Waikato Environment for Knowledge Analysis (WEKA). Its primary function is to extract meaningful information from raw data gathered in agricultural settings. WEKA is designed to advance machine learning methodologies and implement them in solving practical data mining challenges. It directly applies algorithms to datasets and offers support for various standard data mining tasks, including data preprocessing, classification, clustering, regression, visualization, and feature selection. WEKA is a freely available open-source application, governed by the terms of the general public license agreement. It boasts a user-friendly design featuring a graphical interface, facilitating swift installation and ease of use [23]. Over the past few years, WEKA has proven to be effective in handling extensive datasets from various domains such as geology, medicine, marketing, banking, and other business sectors [19].

In Weka, classifiers serve as the main learning techniques, generating rule sets or decision trees to model data. The software encompasses diverse algorithms for acquiring association rules and clustering data. It maintains a consistent command line interface across all applications. The program evaluates the performance of different learning algorithms on a given dataset using a shared evaluation module. Additionally, Weka offers tools or filters for preprocessing data, with standardized command line interfaces akin to learning schemes. Notably, Weka is entirely coded in Java, ensuring the accessibility of data mining tools across various computer platforms. [24].

https://doi.org/10.56038/oprd.v3i1.376

The success of the numerical model to be obtained using the machine learning algorithms (such as Decision tree, Random forest, Logistic regression, K-Nearest neighbor) in the WEKA program is determined by the correlation coefficient (R), coefficient of determination (R^2), mean absolute error (MAE), root mean square error (RMSE), relative absolute error (RAE) and root relative squared error (RRSE) are evaluated according to error values and calculation criteria. Among the evaluation criteria, R^2 is stated to be the accuracy rate and decision-making coefficient of the model [16]. This value varies between 0 and +1, and when this value approaches 0, it means that the model does not adapt to the data [25, 26]. It is said that the model's prediction is better the higher the correlation coefficient [16], a coefficient <0.40 has low correlation, coefficients between 0.40-0.70 have normal correlation, coefficients >0.70 and above have high correlation. It is stated that it indicates good performance results [4, 27, 28, 29]. Since MAE and RMSE, which are one of the most frequently used error measures among estimation methods, are error measures, the performance of the model will be higher if they take low values close to zero [16, 30], and it is better to evaluate them together, but not limited to them. [31], it is stated that the model will give perfect results if MAE [4] and RMSE are equal to zero [32]. Another study states that MAE is a basic accuracy parameter that calculates the average size of the errors of the prediction results, it gives the number differences between the actual and predicted values, and in statistics, the mean absolute error (MAE) is used to measure how close the predictions are to the final results [33]. However, as the distribution of error magnitudes grows more erratic, RMSE tends to be greater than MAE. Whereas RMSE evaluates the average of the mistakes and assigns a relatively high weight to significant errors, MAE measures the average of the errors in a collection of predictions and creates a linear score, indicating that all individual differences are weighted equally on average. The error variance in a series of forecasts may be diagnosed using MAE and RMSE, which have a range of 0 to ∞. [34].

Decision trees are a useful tool for categorizing both numerical and textual data, generally provide a great advantage for decision makers in terms of easy interpretation and understandability compared to other classification techniques [2, 5, 11, 12, 13]. In addition, decision trees are one of the most preferred classification techniques due to reasons such as low cost, faster than other classification techniques, easy to integrate with databases and high reliability [12, 35].

Decision tree applications find utilization in a variety of fields for the purpose of uncovering knowledge and analyzing patterns. These algorithms also offer the opportunity to generate a clear classification/regression model with satisfactory accuracy

https://doi.org/10.56038/oprd.v3i1.376

across diverse domains, including medical diagnosis and credit risk assessment. In this context, decision trees serve as a non-parametric approach employed to enhance classification or regression equations. Regression is represented through hierarchical data structures in supervised learning, wherein independent variables are segmented to forecast the dependent variable. The decision tree has the ability to classify independent variables starting from the root node and provide regression equations with the help of algorithms on each class label. To calculate the value of the dependent variable in a specific scenario, it is necessary to start from the root in the decision tree and follow the edges according to the results of the tests on the attributes. A traditional classification decision tree keeps class labels in its leaves. Figure 2. gives the decision tree, which is a structural representation of how to make a classification starting from the root node according to the X feature. Thanks to this tree, more concise and attractive information can be presented [19, 36].

The decision tree's hierarchical format, typically depicted as a tree with branches and leaves, allows the examination of various factors. The classification outcome is displayed at each leaf, while the branches represent attribute conditions. By employing specific learning strategies on a set of training examples containing input variables and corresponding output variables, a decision tree can be generated to classify variables and establish induction rules. [13; 37].

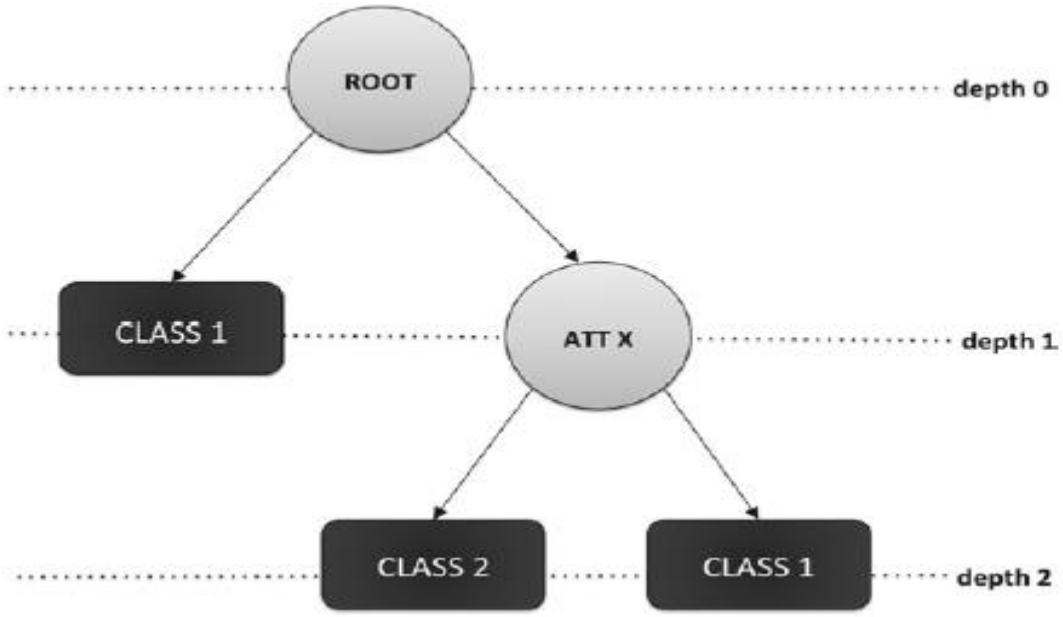

Figure 2. General structure of a decision tree example of the Classification Algorithm [36]

https://doi.org/10.56038/oprd.v3i1.376

Decision trees; They can be listed as algorithms such as DecisionStump, HoeffdingTree, J48, LMT, M5P, RandomForest, RandomTree and REPTree [12, 34, 38]. The selection of decision tree types summarized below varies according to the user's wishes.

a) DecisionStump: consists of a single-level decision tree for numerical or categorical data sets. In the tree created with this algorithm, the root nodes are directly connected to the leaf nodes and the classification process is carried out directly with a single input [4, 12, 19, 39].

b) HoeffdingTree: It is one of the incremental decision tree algorithms. It is a learning algorithm with a tree-building structure from large data sets, assuming that the data sets that produce the distribution do not change over time [4, 19, 40].

c) J48: This algorithm, which is an improved version of the C4.5 algorithm, consists of a binary tree structure instead of multiple tree branches [4]. Developing a decision node using the expected predictions of the class, J48 also deals with estimating missing or missing features of data with certain features and changing feature costs, and its accuracy can be expanded by pruning [41].

d) LMT (Logistic Model Trees): It involves generating trees using the logistic model and follows a conventional decision tree structure, incorporating logistic regression functions at the terminal nodes [19, 42]. This model is widely used in the classification model and can work for attributes with numerical, categorical and missing values [4]. LMT is also a data mining algorithm that combines decision tree (DT) and logistic regression (LR) algorithms [43].

e) RandomForest: These models are made up of regression trees and unpruned raw classification trees that are created by randomly choosing samples from the data sets. The strength of each tree separately and the relationships among these trees determine the generalization error in classification in this model. The approach outperforms Adaboost in terms of results and resilience to noisy values when the characteristics used to break each bond are randomly selected[5, 44]. Both regression and classification (multi-class) can be performed with the algorithm, which provides some advantages to users in terms of calculation and statistics. Data may be trained and predicted rather quickly with just one or two tuning parameters. It is capable of handling high-dimensional issues directly and comes with an integrated generalization error estimate[45].

https://doi.org/10.56038/oprd.v3i1.376

f) RandomTree: is a tree creation method that takes into account a certain number of random features at each node and does not prune [4].

g) REPTree: The algorithm used only in sorting numerical values is fast tree learning using reduced error and is one of the fastest classification algorithms. It uses the information gain criterion in the formation of regression or decision trees and prunes the resulting tree based on the reduced error pruning method. If there are missing values, the partitioning method corresponding to the C4.5 algorithm is used [12, 46].

h) M5P algorithm: It is the Java adaptation of the M5 algorithm developed by John Ross Quinlan on WEKA [47] and derives linear regression inequalities at the nodes of the leaves in the decision tree structure known as M5. The inductive tree algorithm is used to create this tree, and at each stage the independent variables are divided into tree nodes depending on the dependent variable. If the class value at the nodes differs little or the number of nodes is low, the splitting process is stopped, then each leaf of the tree is checked, pruning is performed and regression inequalities are derived from the pruned nodes [37, 38, 48, 49, 50, 51]. There are two separate numerical data in parentheses next to the last leaves of the model obtained using the M5P algorithm; the first is the number of samples representing the leaf, and the second is the mean square error of the predictions from the linear model of each leaf, expressed as a percentage of the standard deviation calculated over all data [4, 38, 52]. The second number in this model is obtained by dividing the mean square error of the linear model in each leaf by the global absolute deviation [53, 54].

2.2. Evaluation of Data

Evaluation, which serves two purposes, is one of the very important key points of the data mining process. The first of these is to estimate how well the final model will work in the future (whether the model will be used or not), and the second is to find the model that best represents the entire data set [19, 55]. In many cases, estimating the model is not as simple as it seems. The main problem here arises when the data used to train a data mining model is used to predict the performance of this model. By storing all training models in the dataset, this evaluation technique can create an unrealistic and overly optimistic perfect estimate that is unacceptable to the data mining community [18, 55]. This method should be used if an explanatory model rather than a predictive model is to be built. Another method that can be used to estimate the model is that if you have a very large data set, the data set can be divided using another program and a separate test set can be used to evaluate the performance of the model. In this way, the entire data set can

https://doi.org/10.56038/oprd.v3i1.376

be used as training data [18]. When limited data is available, more complex methods are required to obtain an unbiased estimate of model performance on unseen samples. These methods; k-fold cross validation, leave-one-out, hold-out and Bootstrap models, which offer alternatives to the user for evaluation depending on the number of available samples. Evaluation of the model actually guides the designer in choosing the best technique [55].

The data classification part consists of learning and classification stages. In learning, training data is analyzed with classification algorithms, while in classification, test data is used to estimate the accuracy of classification rules, and if the accuracy is acceptable, the rules can be applied to new data sets[19, 56]. In order to accurately predict the model result based on the data set, the data is divided into two parts as "training and test" data and stored as a file [55, 57]. Initially, a classifier (decision tree, neural network, etc.) is created with the training set, and then the data in the test set is used to estimate the model with the help of the classifier. If the test set contains C correctly classified data and N samples, the prediction accuracy of the classifier for the test set is p = C/N. This equation can be used to predict performance on any unseen dataset. The prediction model based on the training and test data of the data set is given in Figure 3. To estimate the model, the data set can be divided into training and test data in ratios such as 1:1, 2:1 70:30, 60:40 [57], 66:34 [18] according to the user's purpose. Here, it is generally preferred that the training set consists of as much data as possible in order to obtain a stronger model [18, 57, 58]. A certain amount of the data set (20% - 30%) is kept for testing data, which is called the storage procedure, and then the remaining amount can be used for training. In case the number of data is limited, it is a very common method to allocate one-third of the available data for testing and the remaining two-thirds for training [59]. On the other hand, this method should not be used in cases of small data set and suspicion that the training and test data do not reflect the model.

In general, k-fold cross validation gives better results than the methods listed above [18]. However, this method should not be used if you have very large data. Depending on the size of the dataset, the data can be divided by 5 or 10 and the accuracy of the model can be tested. In practice, the dataset is usually divided into 10, so that 10% of the data is used as test data and 90% as training data in the model [18, 57]. The estimation model of the data set through k-fold cross validation is given in Figure 4 [57].

https://doi.org/10.56038/oprd.v3i1.376

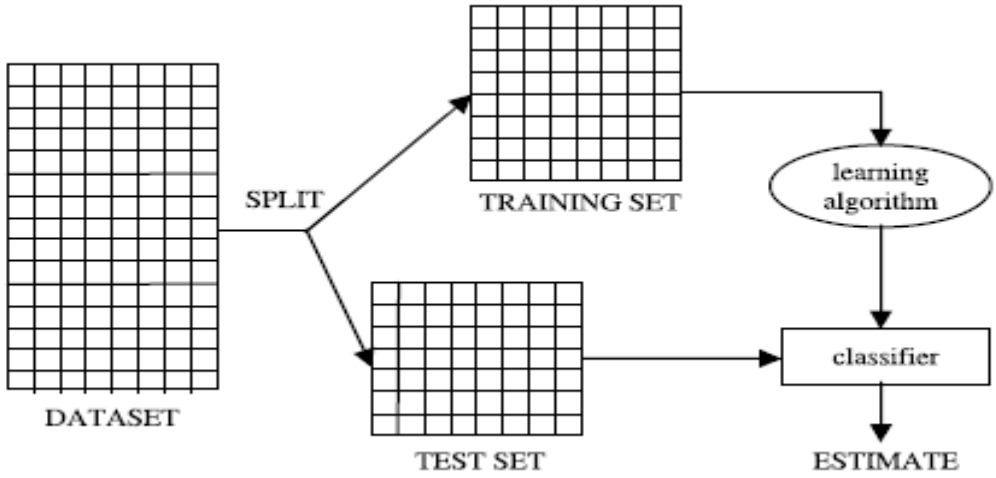

Figure 3. Prediction model of the dataset based on training and test data [57]

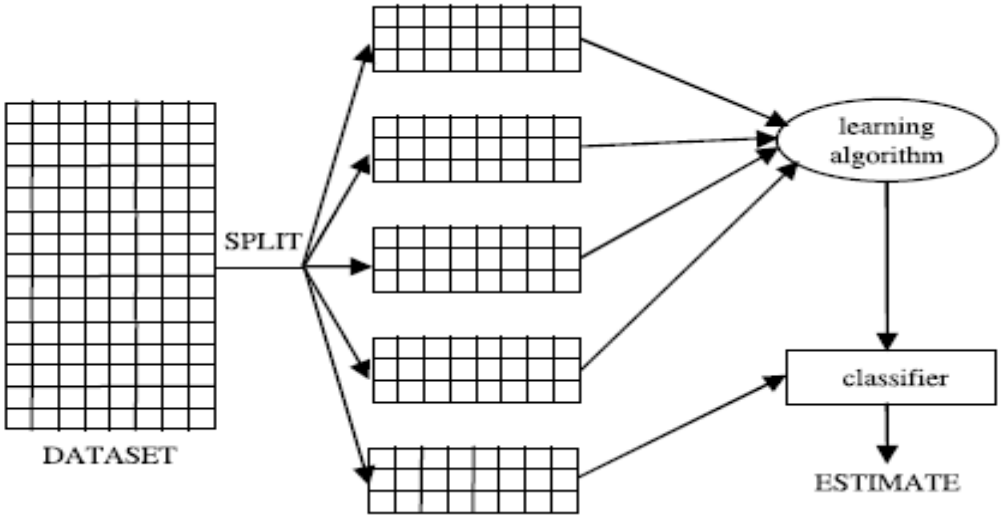

Figure 4. Prediction model of the dataset via k-fold cross validation [57]

2.3. Studies on Decision Trees in Engineering

When the literature is examined, many studies have been carried out using the decision tree algorithm. One of these fields is engineering fields. It is possible to say that the studies on engineering are quite comprehensive. NavieBayes, decision trees and rule finding algorithms, which are among the classification techniques, were used to predict the malfunctions that hinder trips by using data sets of approximately 4 years of tram

https://doi.org/10.56038/oprd.v3i1.376

malfunction records in a tram company in our country. By interpreting the rules obtained as a result of the analysis; It has been stated that the malfunctions that hinder the voyages are concentrated in August and September, and that air temperature, dew point, sea level pressure, rainy and stormy weather and some ferry conditions are effective. On the other hand, it has been stated that cabin condition, humidity rate, wind speed, gender of drivers, fault zones, fault code, equipment name and line information are not effective [60]. Models to predict air pollution have been developed with machine learning algorithms using some meteorological variables in Kastamonu. In prediction models, Artificial Neural Networks (ANN), Random Forest, K-Nearest Neighborhood, Logistic Regression, Decision Tree, Linear Regression and Simple Bayes (Naive Bayes) methods were used. As a result of the study, it was stated that random forest and decision tree machine learning methods showed the highest performance, while the linear regression method showed the worst performance [16].

A study was conducted to estimate the compressive strength of cement mortar containing high volume fly ash used in the construction industry. For this, 450 different cement mortars were analyzed and the results were modeled. According to the modeling results using M5P-tree and Artificial Neural Network (ANN), it was stated that the most important parameters affecting the cement mortar are fly ash, water/binder ratio and curing time. It was also stated that using the M5P-based model, it was concluded that curing time was the most dominant parameter for predicting the compressive strength of cement mortar with this data set [61]. Predictive models of asphalt concretes' dynamic modulus were created using the M5P model tree algorithm. In the study, a data set containing binding properties, grading properties, volumetric properties and test conditions was used to develop prediction models. Consequently, it has been reported that the models created with the M5P algorithm outperform the models created earlier, and the model's performance is greatly enhanced by the logarithmic transformation of the dynamic module values [62].

A decision tree algorithm powered by the sailfish optimization approach, one of the machine learning methods, was used to diagnose heart diseases. It was concluded that accuracy rates improved by 18% on average and the highest improvement was achieved with the decision tree machine learning algorithm of 41.93%. In addition, as a result of the study, it was stated that thanks to the proposed model, an accuracy rate of 0.9836 was obtained, which is also important in the scientific literature for the diagnosis of heart diseases [63]. PM (TAPM, PM10, PM2.5 and PM1), which are released during drilling, loading and transportation activities in a gypsum and two limestone enterprises in Sivas

province and pose a risk to human health, and simultaneously thermal comfort parameters (air temperature, dew) point temperature, side wind, head wind/tail wind, relative humidity, station pressure and wind speed) measurements were carried out. While the measurements were being made, samples representing the properties of the material in working conditions were taken, moisture and silt + clay analyzes were performed in the laboratory and the machine properties were also recorded. PM emission was taken as the dependent variable, and atmospheric conditions, material and equipment properties were taken as the independent variables. As a result, oscillation prediction formulas were derived independently for each PM (TAPM, PM10, PM2.5 and PM1) dimension with the decision tree algorithm, and it was stated that high corrected determination coefficients that could reflect the explanatory capacity of the models were obtained [64].

3. Result

One of the methods used to transform large and complex raw data into useful information through different analysis methods is data mining. Today, the areas where data mining is used are constantly evolving. Computer software is needed to develop and implement data mining. The most important algorithms in the classification techniques used in the analysis of data in these software include Artificial Neural Networks, Support Vector Machines K-Nearest Neighbour, Logistic Regression, Genetic Algorithms and Decision Trees. Among these algorithms, decision tree algorithms are frequently preferred for their easy interpretation and understandability. One of the most important open source data mining programs is WEKA. Thanks to this program, large and complex numerical and alphanumeric data sets can be evaluated with the decision trees algorithm. In this study, different decision tree algorithms in the WEKA program are mentioned and information about the basic functions of each decision tree is given. Thus, it is possible to choose the desired decision tree algorithm according to the user's needs.

References

[1] Albayrak, A. S., Yılmaz, Ş. K. (2009). Veri madenciliği: Karar ağaç algoritmaları ve İMKB verileri üzerine bir uygulama. Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 14(1), 31-52.

[2] Czajkowski, M., Kretowski M. (2010). Globally induced model trees: an evolutionary approach. 11th International Conference on Parallel Problem Solving from Nature, September 11-15, 324-333, Krakow.

[3] Göndör, M., Bresfelean, V. P. (2012). REPTree and M5P for measuring fiscal policy influences on the Romanian capital market during 2003-2010. International Journal of Mathematics and Computers in Simulation, 6(4), 378-386.

[4] Aydemir, E. (2018). Weka ile yapay zekâ. Seçkin Yayınevi, 231s, Ankara.

[5] Onan, A. (2015). Şirket iflaslarının tahmin edilmesinde karar ağacı algoritmalarının karşılaştırmalı başarım analizi. Bilişim Teknolojileri Dergisi, 8(1), 9-19. https//doi.org/10.17671/btd.36087.

[6] Friedman F., Hastie T., Tibshirani R. (2009). The elements of statistical learning data mining, inference and prediction, 2nd Ed., Springer series in Statistics, Springer, 745p, New York.

[7] Küçükönder, H., Vursavuş, K. K., Üçkardeş, F. (2015). K-star, rastgele orman ve karar ağacı (C4.5) sınıflandırma algoritmaları ile domatesin renk olgunluğu üzerinde bazı mekanik özelliklerin etkisinin belirlenmesi. Türk Tarım - Gıda Bilim ve Teknoloji Dergisi, 3(5), 300-306.

[8] Shearer, C. (2000). The Crisp-DM model: the new blueprint for data mining. Journal of Data Warehousing, 5(4), 13-23.

[9] Savaş, S., Topaloğlu, N., Yılmaz, M. (2012). Veri madenciliği ve Türkiye'deki uygulama örnekleri. İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, 11(21), 1-23.

[10] Bramer, M. (2007). Principles of data mining. Springer-Verlag London Ltd., 526p, London.

[11] Gargano, M. L., Raggad, B. G. (1999). Data mining-a powerful information creating tool. OCLC Systems & Services, 15(2), 81-90.

[12] Aydemir, E., Kaysi, F., Yavuz, M. (2020). İlaç satış verileri kullanılarak ağaç algoritmaları ile elde edilen gelirin tahmin edilmesi. Anatolian Journal of Computer Sciences, 5(1), 14-21.

[13] Chien, C. F., Chen, L. F., (2008). Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry, Expert Systems with Applications, 34(1), 280-290.

[14] Albayrak, A. S., Yılmaz, Ş. K. (2009). Veri madenciliği: Karar ağaç algoritmaları ve İMKB verileri üzerine bir uygulama. Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 14(1), 31-52.

[15] Gorunescu, F. (2011). Data mining: concepts, models and techniques. Springer-Verlag, 370p, Heidelberg.

[16] Gültepe, Y. (2019). Makine öğrenmesi algoritmaları ile hava kirliliği tahmini üzerine karşılaştırmalı bir değerlendirme. European Journal of Science and Technology, 16, 8-15.

[17] Chadha, P., Singh, G. N. (2012). Classification rules and genetic algorithm in data mining. Global Journal of Computer Science and Technology Software & Data Engineering, 12(15), 50-54.

[18] Brownlee, J. (2016). Machine learning mastery with Weka, Machine Learning Mastery, 248p.

[19] Aksu, G. (2018). Pisa başarısını tahmin etmede kullanılan veri madenciliği yöntemlerinin incelenmesi. Hacettepe Üniversitesi Eğitim Bilimleri Enstitüsü (Doktora Tezi), 162s, Ankara.

[20] Saygılı, A., (2013). Veri madenciliği ile mühendislik fakültesi öğrencilerinin okul başarılarının analizi. Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü (Yüksek Lisans Tezi), 129s, İstanbul.

[21] Bruxella, J.M. D., Sadhana, S., Geetha, S. (2014). Categorization of data mining tools based on their types. International Journal of Computer Science and Mobile Computing, 3(3), 445-452.

[22] Jović, A., Brkić, K., Bogunović, N. (2014). An overview of free software tools for general data mining. 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), May 26-30, 1-6, Opatija.

[23] Kiranmai, S. A., Jaya Laxmi, A. J. (2018). Data mining for classification of power quality problems using WEKA and the effect of attributes on classification accuracy. Protection and Control of Modern Power Systems, 3, 1-12. https://doi.org/10.1186/s41601-018-0103-3.

[24] Alfred, R., (2005). Knowledge discovery: enhancing data mining and decision support integration. The University of York (Qualifying Dissertation), 45p, York.

[25] Alpar R. (2011). Uygulamalı çok değişkenli istatistiksel yöntemler. Detay Yayıncılık, 853s, Ankara.

[26] Çınaroğlu, S. (2016). Sağlık harcamasının tahmininde klasik regresyon yöntemleri ile veri madenciliği regresyon yöntemlerinin karşılaştırılması. Ekonomik Yaklaşım, 27(101), 185-218.

[27] Schober, P., Boer, C., Schwarte, L. A. (2018). Correlation coefficients: appropriate use and interpretation, Anesthesia & Analgesia, 126(5), 1763-1768. https://doi.org/10.1213/ANE.0000000000002864.

[28] Sabti, A. A., Rashid, S. M., Hummadi, A. S. (2019). Interrelationships between writing anxiety dimensions and writing goal orientation among Iraqi EFL undergraduates, International Journal of Instruction, 12(4), 529-544, https://doi.org/10.29333/iji.2019.12434a.

[29] Tanni, S. E., Patino, C. M., Ferreira, J. C. (2020). Correlation vs. regression in association studies. Jornal Brasileiro de Pneumologia, 46(1): e20200030. https://doi.org/10.1590/1806-3713/e20200030.

[30] Wang, W., Xu, Z. (2004). A heuristic training for support vector regression. Neurocomputing, 61: 259-275. https://doi.org/10.1016/j.neucom.2003.11.012.

[31] Chai, T., Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? - Arguments against avoiding RMSE in the literature. Geoscientific Model Development, 7, 1247-1250.

[32] Çınaroğlu, S. (2017). Sağlık harcamasının tahmininde makine öğrenmesi regresyon yöntemlerinin karşılaştırılması. Uludağ Üniversitesi Mühendislik Fakültesi Dergisi, 22(2), 179-200.

[33] Usha, T. M., Balamurugan, S. A. (2016). Seasonal based electricity demand forecasting using time series analysis. Circuits and Systems, 7(10), http://dx.doi.org/10.4236/cs.2016.710283.

[34] Alsultanny, Y.A. (2020). Machine learning by data mining REPTree and M5P for predicating novel information for PM10. Cloud Computing and Data Science, 40-48.

[35] Akçetin, E., Çelik, U. (2014). İstenmeyen elektronik posta (spam) tespitinde karar ağacı algoritmalarının performans kıyaslaması. İnternet Uygulamaları ve Yönetimi Dergisi, 5(2), 43-56. https://doi.org/10.5505/iuyd.2014.43531.

[36] Barros, R. C., de Carvalho, C. P. L. F. A., Freitas, A.A. (2015). Automatic design of decision-tree induction algorithms. SpringerBriefs in Computer Science, 176p, London.

[37] Njeguš, A., Vanja Nikolić, V., Jovanović, V. (2015). The selection of optimal data mining method for small-sized hotels. International Scientific Conference of IT and Business-Related Research, April 16, 519-524, Belgrade.

[38] Witten, I. H., Frank, E., Hall, M. A. (2011). Data mining: practical machine learning tools and techniques. Morgan Kaufmann Publishers, 665p, Burlington. https://doi.org/10.1016/C2009-0-19715-5.

[39] Shah, T. N., Khan, M. Z., Ali, M., Khan, B., Idress, N. (2020). CART, J-48graft, J48, ID3, decision stump and random forest: a comparative study. University of Swabi Journal, 2(1), 1-6.

[40] Srimani, P. K., Patil, M. M. (2015). Performance analysis of Hoeffding trees in data streams by using massive online analysis framework. International Journal of Data Mining Modelling and Management, 7(4), 293-313. http://dx.doi.org/10.1504/IJDMMM.2015.073865.

[41] Saravanan, N., Gayathri, V. (2018). Performance and classification evaluation of J48 algorithm and Kendall's based J48 algorithm (KNJ48). International Journal of Computer Trends and Technology, 59(2), 188-198. https://doi.org/10.14445/22312803/IJCTT-V59P112.

[42] Landwehr, N. (2003). Logistic model trees. Computer Science at the University of Freiburg (Diploma Thesis), Germany, 104p, Freiburg.

[43] Maulana, M. F., Defriani, M. (2020). Logistic model tree and decision tree J48 algorithms for predicting the length of study period, Journal Penelitian Ilmu Komputer, System Embedded & Logic, 8(1), 39-48. https://doi.org/10.33558/piksel.v8i1.2018.

[44] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324.

[45] Cutler, A., Cutler, D. R., Stevens, J. R. (2011). Random forests, Machine Learning, 45(1), 157-176. doi: 10.1007/978-1-4419-9326-7_5.

[46] Zhao, Y., Zhang, Y. (2008). Comparison of decision tree methods for finding active objects. Advances in Space Research, 41(12), 1955-1959. https://doi.org/10.1016/j.asr.2007.07.020.

[47] Quinlan J.R. (1992). Learning with continuous classes. 5th Australian Joint Conference on Artificial Intelligence, 343-348, Singapore.

[48] del Campo-Avila J., Moreno-Vergara N., Trella-Lopez M. (2011). Analyzing factors to increase the influence of a Twitter user. Advances in Intelligent and Soft Computing, 89, 69-76.

[49] Öztürk, E. (2012). Görüntü sıkıştırma yöntemlerinin etkinliğini arttıran dönüşüm ve bölümlendirme işlemleri. Trakya Üniversitesi Fen Bilimleri Enstitüsü (Yüksek Lisans Tezi), 84 s, Edirne.

[50] Kara, Ş. E., Şamlı, R. (2021). Yazılım projelerinin maliyet tahmini için WEKA'da makine öğrenmesi algoritmalarının karşılaştırmalı analizi. Avrupa Bilim ve Teknoloji Dergisi, 23, 415-426. doi: 10.31590/ejosat.877296.

[51] Sihag, P., Singh, B., Said, A., Azamathulla, H. M. (2021). Prediction of Manning's coefficient of roughness for high-gradient streams using M5P. Water Supply, 22(3), 2707-2720. https://doi.org/10.2166/ws.2021.440.

[52] Url-1 <https://stats.stackexchange.com/questions/228724/m5p-interpretations-and-questions> alındığı tarih: 20.05.2022.

[53] Url-2 <https://community.rapidminer.com/discussion/440/the-regression-trees-returned-by-the-operators-w-m5p-and-w-reptree> alındığı tarih: 20.05.2022.

[54]Url-3 <https://list.waikato.ac.nz/hyperkitty/list/wekalist@list.waikato.ac.nz/thread/AA5GPEFMQHXXDT6G4HCINHY52UHODW3Z> alındığı tarih: 20.05.2022.

[55] Souza, J., Matwin, S., Japkowicz, N. (2002). Evaluating data mining models: a pattern language. 9th Conference on Pattern Language of Programs (PLOP'02), September 8-12, Monticello.

[56] Ramageri, M. B. (2010). Data mining techniques and applications. Indian Journal of Computer Science and Engineering, 1(4), 301-305.

[57] Bramer, M. (2013). Principles of data mining (2nd ed.), Springer-Verlag, 455p, London.

[58] Genç, B., Tunç, H. (2019). Optimal training and test sets design for machine learning, Turkish Journal of Electrical Engineering & Computer Sciences, 27, 1-13. doi:10.3906/elk-1807-212.

[59] Aksu, G., Doğan, N. (2019). An analysis program used in data mining: WEKA. Journal of Measurement and Evaluation in Education and Psychology, 10(1), 80-95.

[60] Turna, F., (2011). Veri Madenciliği Teknikleriyle Tramvay Arıza Kayıtlarından Kural Çıkarımı, Erciyes Üniversitesi, Fen Bilimleri Enstitüsü, Endüstri Mühendisliği Anabilim Dalı (Yüksek Lisans Tezi), 89 s, Kayseri.

[61] Mohammed, A., Rafiq, S., Sihag, P., Kurda, R., Mahmood, W., Ghafor, K., Sarwar, W., (2020). ANN, M5P-tree and nonlinear regression approaches with statistical evaluations to predict the compressive strength of cement-based mortar modified with fly ash, Journal of Materials Research and Technology, 9(6):12416-12427. https://doi.org/10.1016/j.jmrt.2020.08.083

[62] Behnood, A., Daneshvar, D., (2020). A machine learning study of the dynamic modulus of asphalt concretes: An application of M5P model tree algorithm, Construction and Building Materials 262, 120544, https://doi.org/10.1016/j.conbuildmat.2020.120544

[63] Yıldırım, M, O., (2021). Yelken Balığı Eniyileme Yaklaşımı ile Güçlendirilmiş Karar Ağacı Algoritması Kullanarak Kalp Rahatsızlıklarının Teşhisi, Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Endüstri Mühendisliği Anabilim Dalı (Yüksek Lisans Tezi), 64 s, Isparta.

[64] Duran, Z., (2022). Bazı açık maden işletmelerinde partikül madde salınım ölçümü ve değişiminin meteorolojik koşullar, malzeme ve iş makinesi özellikleri ile modellenmesi, Sivas Cumhuriyet Üniversitesi Fen Bilimleri Enstitüsü Maden Mühendisliği Ana Bilim Dalı (Doktora Tezi), 380 s, Sivas.