Makalenin Geliş Tarihi: 27.08.2021
1. Hakem Rapor Tarihi: 04.10.2021
2. Hakem Rapor Tarihi: 02.11.2021
Kabul Tarihi: 02.11.2021

A REVIEW ON DATA MINING METHODS USED IN INTERNAL AUDIT AND EXTERNAL AUDIT [*]

(Araştırma Makalesi)

Jale SAĞLAR [**] - İlker KEFE [***]

Abstract

In this study, data mining methods used in audit activities are explained. Based on the results of the research on data mining, common data mining methods have been determined and the usability of these methods in audit activities is examined. In addition, the analyzed data mining methods were discussed in terms of fraud detection and the cost created by fraud. The study also evaluates which data mining method or methods are more appropriate to prevent these costs. This study focuses on DM techniques, especially artificial neural networks (ANN), logistic regression (LR), decision trees (DT), support vector machines (SVM), genetic algorithms (GA), and text mining (TM).

Keywords: *Data Mining, Internal Auditing, External Auditing, Fraud, Fraudulent Financial Reporting.*

Jel Codes: *M42, C15, C45*

İç Denetim ve Bağımsız Denetimde Kullanılan Veri Madenciliği Yöntemleri Üzerine Bir İnceleme

Öz

Bu çalışmada denetim faaliyetlerinde kullanılan veri madenciliği yöntemleri incelenmektedir. Veri madenciliği ile ilgili araştırma sonuçlarına dayalı olarak yaygın veri madenciliği yöntemleri belirlenmiş ve bu yöntemlerin denetim faaliyetlerinde kullanılabilirliği incelenmiştir. Ayrıca analiz edilen veri madenciliği yöntemleri, hile

[*] Bu makale, 23-25 Nisan 2021 tarihleri arasında gerçekleştirilen VI. Uluslararası Muhasebe ve Finans Sempozyumunda sunulmuş bildirinin genişletilmiş hâlidir. Çalışma Ç.Ü. Bilimsel Araştırma Projeleri Birimi (Proje No: 12667) tarafından desteklenmektedir.

[**] Doç. Dr., Çukurova Üniversitesi, İİBF İşletme Bölümü, İşletme Anabilim Dalı (e-posta: jsaglar@cu.edu.tr). ORCID ID: https://orcid.org/0000-0001-7152-9807

[***] Dr. Öğr. Üyesi, Osmaniye Korkut Ata Üniversitesi, İİBF Uluslararası Ticaret ve Lojistik Bölümü, İşletme Anabilim Dalı, (e-posta: ilkerkefe@osmaniye.edu.tr) ORCID ID: https://orcid.org/0000-0002-9945-5325

tespiti ve hilenin yarattığı maliyet açısından ele alınmıştır. Çalışmada bu maliyetleri önlemek için hangi veri madenciliği yöntem veya yöntemlerinin daha uygun olduğu değerlendirilmektedir. Çalışmada veri madenciliği yöntemlerinden özellikle Yapay Sinir Ağları, Lojistik Regresyon, Karar Ağaçları, Destek Vektör Makineleri, Genetik Algoritmalar ve Metin Madenciliği tekniklerine odaklanmaktadır.

Anahtar Kelimeler: *Veri Madenciliği, İç Denetim, Bağımsız Denetim, Hile, Hileli Finansal Raporlama.*

Jel Kodlar: *M42, C15, C45*

1. Introduction

Today, there is a large number of digital data formation in the electronic environment. Data are produced, stored, and evaluated in the digital environment. In addition, the desired results are expected to be achieved quickly. Traditional tools are not enough to analyze big data. The main reason for this is that the data volume is large, it evolves very quickly, and its variability or relevance has changed significantly over time (Gepp, Linnenluecke, O'Neill, and Smith, 2018).

Auditors are responsible for detecting fraudulent transactions and preventing their recurrence (Holton, 2009). Auditors should use more up-to-date and innovative techniques in the detection of financial statement fraud in order to prevent the negative consequences of fraud (Ata and Seyrek, 2009). Data mining (DM) techniques have increased as a result of the importance of big data. Training and investment activities are needed for the learning and practical use of data mining techniques (Hassani, Huang, Silva, and Ghodsi, 2016). The American Institute of Certified Public Accountants (AICPA, 1999) states that data mining is one of the ten best technologies of the future (Koh and Low, 2004).

DM is one of the key components of advanced intelligent business analytics and decision support tools (Amani and Fadlalla, 2017). The use of DM in academic studies, medical and scientific research has increased significantly. DM applications are also effectively applied in many different areas such as manufacturing and service industries, retail, banking, insurance, tourism, and telecommunications (Xiao, Xiaoli, and Gaojin, 2010). DM has recently gained interest and increased popularity in the field of finance and accounting (Zhou and Kapoor, 2011). The main focus of DM is to use the data assets of the business to obtain useful financial and non-financial information. Therefore, it can be said that DM is applied to many business disciplines, including accounting (Amani and Fadlalla, 2017).

Financial statement fraud has negative effects on individual investors and more importantly on the overall stability of global economies (Zhou and Kapoor, 2011). With the detection of financial fraud, possible destructive effects of financial fraud are prevented. Financial fraud detection primarily focuses on the detection of fraudulent financial data, the visibility of fraudulent behavior and activities, and the development of appropriate strategies by decision makers (Ngai, Hu, Wong, Chen, and Sun, 2011). Creating an opportunity to commit fraud is the sore point of fraud detection software. A typical procurement control tool should have access to all relevant information and documents. All proving information and documents such as invoice, order form, order slip, payment amount must be ready in order to determine that the correct accounting process for the purchased goods and services has been made. Duplicate payments, payments that exceed authorization levels, or payments made on weekends may be specific targets for this analysis activity. Audit tools flag unusual values for investigation and sometimes add a risk score to each identified potential fraud indicator (Holton, 2009). When evaluated in terms of audit activities, big data techniques make a great contribution to the audit, regardless of whether the client company uses it or not (Gepp et al. 2018). However, DM techniques have not yet been systematically integrated into accounting and the potential benefits of DM techniques to accounting have not yet been fully disclosed (Amani and Fadlalla, 2017).

In this study, data mining methods used in audit activities are mentioned. In this context, data mining methods are investigated by focusing on the studies in the field of accounting in data mining. In the study, fraud detection of data mining methods, the cost created by fraud and which methods are more effective to prevent these costs are evaluated.

Ethics of Research

This study was conducted according to ethical considerations. Scientific ethical rules were taken into consideration in the study. The citations in the study are shown accurately and completely. The study has not been submitted for evaluation to another academic publication. In addition, the authors confirm that the study does not require ethics committee approval.

2. Importance of Data Mining

DM is a technique where ideas from different disciplines and fields of study such as statistics, electronics, computers, artificial intelligence, modeling, and information technology can be used integrated (Xiao et al. 2010). DM can be expressed as a multidisciplinary approach using many different techniques (Amani and Fadlalla, 2017). DM refers to obtaining the desired information from a large data stack (Han and Kamber, 2006). DM aims to discover hidden connections between data and focuses on making interrelated connections in big data (Keyvanpour, Javideh, and Ebrahimi, 2011). DM conducts searches within one or more data networks. The aim of the results obtained is an estimate of possible future outcomes (Gray and Debreceny, 2014). DM methods allow for the efficient processing of data with missing values, irrelevant data or highly correlated datasets automatically. These methods become important, especially during real-time data flow, where manual data entry is not possible or difficult (Gepp et al. 2018).

DM techniques are used by businesses for many different purposes. Businesses utilize DM techniques to identify previously undetectable relationships or connections among their data, to establish predictive analytics and to create new types of performance metrics (Yigitbasioglu and Velcu, 2012). DM techniques can make predictions about the future corporate development trend of a business, help senior management in the decision-making process, and provide results that will increase the competitiveness of the business (Xiao et al. 2010).

A wide variety of DM techniques are available for data analysis such as artificial neural networks (ANN), association rules, case-based reasoning, decision trees (DT), fuzzy analysis, genetic algorithms (GA), k-nearest neighbor, logistic regression (LR), naïve Bayes, random forest, self-organizing maps, support vector machines (SVM), and text mining (TM) (Albizri, Appelbaum, and Rizzotto, 2019; Amani and Fadlalla, 2017; Bhattacharyya, Jha, Tharakunnel, and Westland, 2011; Bauder, Khoshgoftaar, and Seliya, 2017; Koh and Low, 2004; Lin, Chiu, Huang, and Yen, 2015).

Koh and Low (2004) investigated and compared predictive modelling techniques in data mining. Each of the DM techniques helps with a specific purpose, problem-solving, and business need (Amani and Fadlalla, 2017). In this study, DM techniques ANN, LR, DT, SVM, GA and TM have been included. In this context, each method is defined and how they are used in audit activities are explained. Next, data mining techniques commonly used in audit activities will be explained.

2.1. Artificial Neural Networks (ANN)

Neural networks, also called artificial neural networks, are designed to simulate the human brain (An, 2009). ANN is a method developed to have the ability to generate and discover new information through self-automatic learning, similar to the functioning of the human brain. ANN is an information processing system that simulates human brain functions such as thinking and learning (Yildiz and Yezegel, 2010). ANN has similar features to the structure of biological neural networks in the brain. This structure is similar to the working principle of neurons in our brain (Ata and Seyrek, 2009). In fact, the human brain has a much more complex structure than an ANN developed so far (An, 2009). ANN is actually a supervised machine learning algorithm (Dutta, Dutta, and Raahemii, 2017) and nonparametric model designed by considering the learning processes of a brain (Gepp et al. 2018). An artificial neural network needs many interconnected input streams for input and output generation (Ata and Seyrek, 2009). Each unit created in ANN actually computes a simple function. Similar to the human brain, an artificial neural network consists of interconnected neurons and node units (An, 2009). ANN imitates learning methods, the ability to remember and generalize when necessary. ANN also consists of units or neurons that mimic a biological neural network. These units and neurons form many simple processors and carry digital data through communication channels (Rivero, Rabuñal, Dorado, and Pazos, 2009). ANN includes training examples and input-output training pairs. Thus, information about the transformation to be realized by providing iterative learning is obtained (Lin et al. 2015). Here, ANN is trained by teaching the patterns of the network and the relationships between each other with the help of sample inputs (Ata and Seyrek, 2009). The network is actually trained to find the correct answer or learn the outcome for each of the training examples (Feroz, Kwon, Pastena, and Park, 2000). Then the results are learned and can be generalized. Thus, the results learned can be used on new data (Ata and Seyrek, 2009). In summary, ANN is useful for designing nonlinear systems using a large number of inputs. The design is based only on examples of input-output relationships (Lin et al. 2015).

2.2. Logistic Regression (LR)

Regression is a statistical method that detects whether there is a relationship between one or more independent variables and a dependent variable (Han and Kamber, 2006). Regression focuses on the prediction of the dependent variable based on the independent variable (Amani and Fadlalla, 2017). The main purpose of LR is to model the link between a dependent variable and some independent variable (Dalla Valle, 2009). LR is one of the traditional statistical methods (Koh and Low, 2004) and stands out among DM applications used in real life (Bhattacharyya et al. 2011). LR is a linear discrimination classification. LR is used to explain the relationship between a dependent binary variable and one or more independent variables. The metrics in the independent variable can be interval or ratio scale (Hajek and Henriques, 2017). Traditional regression models need to be more flexible. Statistical techniques can be used more effectively if a suitable structure is created by considering the diversity and speed of the data volume (Gepp et al. 2018).

2.3. Decision Trees (DT)

DT are hierarchical or tree-structured forecasting models and are most used in classification and forecasting methods (Lin et al. 2015). DT have flowcharts that look like a tree structure (Hajek and Henriques, 2017). DT are created by many nodes and branches at different stages and under various conditions (Lin et al. 2015). Essentially, the purpose of the tree is to classify data according to the discrete values of a target variable using several predictive variables (Ata and Seyrek, 2009). In this structure, nodes represent attributes and branches represent possible attribute values (Hajek and Henriques, 2017). The nodes of a decision tree represent the test points of the prediction variables. Depending on the result of testing at a node, the tree may be split into more decision nodes or leaf nodes at lower levels (Ata and Seyrek, 2009). As a step of the DT algorithm, the data is first tested. In this way, it is aimed to divide the data into homogeneous subgroups (Siciliano and Conversano, 2009). DT has widely used DM algorithms for classification of problems (Ata and Seyrek, 2009). The purpose of DT is to classify observations by dividing them into mutually exclusive and comprehensive subgroups. This process is carried out by choosing the qualities that best distinguish the sample (Zhou and Kapoor, 2011).

DT are multi-stage decision systems and have a process in which classes are rejected sequentially until one finally arrives at an accepted class (Lin et al. 2015). DT are hierarchical decision models and are generally easy to interpret compared to other methods (Ata and Seyrek, 2009). DT is preferred by businesses due to its advantages such as ease of use, flexible classification of various data and easy interpretation (Bhattacharyya et al. 2011). The most important advantage of DT is that the rules obtained with the help of the model are interpretable (Hajek and Henriques, 2017). DT ensures that the data representing the information becomes meaningful and thus IF-THEN classification rules are determined more easily (Lin et al. 2015).

2.4. Support Vector Machines (SVM)

SVM is one of the statistical learning techniques that can achieve successful results in various classification tasks (Bhattacharyya et al. 2011). SVM is a linear classifier that finds the hyperplane that separates by the largest possible margin, created after converting the data to a high dimension. The largest margin is expressed as the distance between the hyperplane and the nearest data point (Yom-Tov, 2003). SVM aims to correctly classify samples within a dataset. An optimal platform is created that will bring the original space to a higher-dimensional space. Thus, accurate results are achieved by maximizing the margin between samples in the dataset and minimizing the possibility of error (Chrysostomou, Lee, Chen, and Liu, 2009). The purpose of SVM is to specify a classifier or regression function that minimizes empirical risk (ie, training set error) and confidence interval (corresponding to generalization or test set error) (Kurban, Niyaz, and Yıldırım, 2016). SVM is one of the techniques developed for the multidimensional function approximation (An, 2009). SVM uses an optimization criterion based on a compromise between the training error and the complexity of the resulting learning machine (Camps-Valls, Martínez-Ramón, Rojo-Álvarez, 2009). SVM simultaneously maximizes the geometric margin and minimizes the empirical classification error. Therefore, SVM is the maximum margin classifier (Lewis and Ras, 2009). SVM has good theoretical and experimental generalization properties and works well on high-dimensional datasets (Dominik, Walczak, and Wojciechowski, 2009).

2.5. Genetic Algorithms (GA)

GA is one of the problem-solving techniques (Goldberg, 1989). GA contributes to the solution of problems where the number of input features is excessively large (Yom-Tov, 2003). GA is applied in data warehousing and mining. The parallelizable and flexible mechanism of GA enables the search and optimization of many important problems. The GA mechanism includes feature selection, partitioning, and extraction (especially through clustering) (Hsu, 2009). GA uses biologically derived techniques such as heredity, mutation, natural selection, and recombination in the medical world. These techniques are actually in a certain class of evolutionary algorithms. GA is designed as a computer simulation. Abstract representations (chromosomes) of alternatives (individuals) are designed for the solution of an optimization problem and it is aimed to develop a population towards better solutions (Hsu, 2009). In a standard GA, each chromosome represents a possible solution to the problem. The set of all chromosomes is usually called the population. The chromosome can be represented as a binary sequence or as the actual code, depending on the nature of the problem. (Saxena, Kothari, and Pandey, 2009). Traditionally, solutions are represented in binary as sequences of 0's and 1's, but different encodings are possible (Hsu, 2009). The problem is related to a fitness function. After each iteration, the population goes through a series of iterations with crossover and mutation to find a better solution (better fitness value). After a certain fitness level or a certain number of iterations, the process is stopped and the chromosome giving the best fitness value is preserved as the best solution to the problem (Saxena et al. 2009). GA encodes potential solutions to an optimization problem as a chromosome-like data structure and applies recombination operators to these structures. These recombination operators are designed to gradually improve solutions, just as evolution improves individuals in a population (Yom-Tov, 2003).

2.6. Text Mining (TM)

TM is an exploratory data analysis process applied to data in text format (Cerchiello, 2009). TM emerged as a result of the need to use text-based unstructured data (Peji´c Bach, Krstic, Seljan, and Turulja, 2019). TM is used to reveal the data and transform it into meaningful information to be used in the decision-making process (Peji´c Bach et al. 2019). The main purpose of the TM method is to discover unknown information or to find answers to questions (Cerchiello, 2009).

TM is the process of obtaining information by processing unstructured text (Zhu, 2009). Data subject textual documents consist of reports, official documents, plain texts, reviews, e-mails, and web pages (Peji´c Bach et al. 2019). TM implementation process consists of parsing, filtering, categorizing, clustering, and analyzing the text to extract its relevance, usefulness, interestingness, and novelty (Lin and Lehto, 2009). In the TM method, the text is processed in two stages. In the first stage, the document is defined by the categorization process and its content is determined. In the second stage, the document is divided into descriptive categories with the classification process and the relationship between the documents is established (Zhu, 2009). The important point in TM is to organize and structure the text appropriately for further qualitative and quantitative analysis (Peji´c Bach et al. 2019).

3. Data Mining Methods in Financial Fraud Detection and Audit Process

A fraudulent financial statement contains intentionally created false information. The increase in the number of financial statements prepared and published in this way means a serious economic and social problem signal (Zhou and Kapoor, 2011). Access to frequently updated large datasets is critical to audit activities. Updated standards and accessibility have allowed the audit profession to benefit from DM techniques (Gepp et al. 2018). DM techniques used in many fields are also preferred in audit planning activities in recent years (González and Velásquez, 2013). DM methods are used for financial distress forecasting and financial fraud detection in the audit field (Gepp et al. 2018). DM methods used in the audit field are mostly focused on prediction. DM methods are also used in risk management applications such as risk prevention, incident detection, incident mitigation (Amani and Fadlalla, 2017). When DM is used to reveal hidden truths contained in huge amounts of data, it fills an important gap in financial fraud detection (Ngai et al. 2011). DM methods are also frequently used for auditing financial statements (Gray and Debreceny, 2014). Assurance and compliance-oriented DM practices primarily focus on three main topics. These are auditing, business health, and forensic accounting. Auditing mainly includes engagement, planning, conducting, and post-auditing phases. Business health comprises financial viability, bankruptcy, and going concern. Forensic accounting includes fraud detection and earnings management (Amani and Fadlalla, 2017). In some cases, it may be difficult to obtain the desired audit data. Therefore, DM methods should be discovered using unlabeled records for the detection of upcoding fraud (Bauder et al. 2017). Gray and Debreceny (2014) emphasized four points for the use of DM techniques in fraud detection. These are analysis of fraud risks, identification of possible fraudulent methods, determination of fraud indicators and fraudulent methods used, selection of appropriate DM methods most likely to detect these indicators. The use of DM for financial audits is a management decision that requires consideration of its cost to the business. Because the DM method used has many effects that cause costs such as software supply to the company, technological infrastructure formation, personnel training (Gray and Debreceny, 2014). With the widespread use of DM methods, continuous effective use of the data may be possible. By using the available data as a cache, fraudulent transactions and fraud can be detected and costs can be reduced. DM methods have the potential to significantly reduce costs (Bauder et al. 2017). However, additional audit costs may arise from performing unnecessary audit procedures. Here, the auditor should determine the monetary threshold by performing a cost-benefit analysis (Alden, Bryan, Lessley, and Tripathy, 2012). For this reason, DM methods make a positive contribution to the reduction of unnecessary costs for auditors in their reviews.

There are various DM studies on fraud detection, auditing financial statements, and conducting the auditing profession more effectively. Chen, Liou, Chen, and Wu (2019) highlight that several DM approaches have been developed to detect fraud in financial statements. Dutta et al. (2017) characterized widely used DM techniques to detect fraudulently generated financial restatements in their research. Amani and Fadlalla (2017) showed the most widely used DM applications in accounting in their study. Gray and Debreceny (2014) focused on fraud detection with financial statement auditing using DM methods. In the study, a classification is proposed to guide the audit activities, and the quantitative data mainly consisted of financial statements and journal entry data. Yu, Guang, and Zi-qi (2013) developed models based on DM techniques and enabled the detection and classification of violations regarding the disclosure of accounting information of listed companies. Kim and Vasarhelyi (2012) mentioned DM methods used to detect internal fraud at the company level. Goel and Gangolly (2012) examined qualitative text content in annual financial reports to predict fraud. As a result of the study, it has been determined that textual information offers significant clues about fraud prediction. Perols (2011) discusses possible problems that arise with the detection of fraud in financial statements. In the study, the cost differences caused by the wrongly written positive and negative values and the nature of the scattered financial statement data are discussed. Jans, Van Der, Jan, Lybaert, and Vanhoof (2011) examined the effectiveness of audit procedures prepared for detecting fraud during stock and warehousing movements. Debreceny and Gray (2011) presented a framework for how DM techniques can be applied to e-mails and how auditors can use the data obtained as a result of DM as audit evidence. Jans, Lybaert, and Vanhoof (2010) made clustering using DM methods to identify possible fraudulent transactions in the procurement process in a company. The main purpose of the study is to detect and reduce the risk of internal fraud in a large data set. Gaganis (2009) used DM classification techniques combining financial and non-financial data to identify fraudulent financial statements. Kirkos, Spathis, and Manolopoulos (2007) analyzed the effectiveness of DM methods to identify the nature of fraudulent financial statements and factors associated with fraudulent transactions in financial statements. Busta and Weinberg (1998) focused on detecting manipulated fraudulent financial data in their study.

DM methods are used for financial statement audits, financial distress prediction, financial fraud detection, and have different working principles and simulations. For this reason, the following section explains which DM methods are used in detecting fraudulent transactions and how they are used in audit activities. Ağdeniz and Yıldız (2018) emphasized the abundance of studies on financial statement fraud in studies in the field of auditing on TM. Lin, Chiu, Huang, and Yen (2015) demonstrated DM techniques using LR, ANN, and DT to detect financial statement fraud. González and Velásquez (2013) used the DT method with various variables. These variables were used to determine the behaviors towards fraud or not, to identify interrelated behavior patterns, and to measure to what extent fraud could or could not be detected with the data available. Gupta and Gill (2012) suggested using SVM and TM to detect fraud in qualitative parts of financial statements. Alden et al. (2012) used GA to detect and classify fraudulent financial statement patterns. Zhou and Kapoor (2011) discussed DM methods such as ANN, Bayesian networks, DT, and LR that provide financial fraud identification. Bhattacharyya et al. (2011) investigated the performance of LR and SVM techniques in predicting fraud. Ravisankar, Ravi, Rao, and Bose (2011) used DM techniques such as genetic programming, group method of data handling, LR, multilayer feed forward neural network, probabilistic neural network, and SVM to identify companies that manipulate financial statements. Ngai et al. (2011) suggested DM methods such as ANN, DT, LR, and the Bayesian belief network for the detection and classification of fraudulent data. In the study, solution suggestions are also presented for possible problems while performing data analysis. Liou (2008) investigated the differences and similarities between fraudulent financial reporting detection and business failure prediction models and found that LR and DT gave the best results among DM methods in fraudulent reporting. Welch, Reeves, and Welch (1998) discussed the classification problem involving supervisors' decisions and proposed a DM-based GA method for classification decision models.

In terms of internal auditing and external auditing, some data mining methods become prominent. Table 1 shows which data mining techniques can be used in the internal audit and external audit process.

Table 1. Data Mining Methods Used in The Internal and External Audit Process

		METHODS					
		Artificial Neural Networks	***Logistic Regression***	***Decision Trees***	***Support Vector Machines***	***Genetic Algorithms***	***Text Mining***
AUDITING PROCESS	***Internal Auditing Process***	[1], [2], [9], [23]	[3], [23]	[4], [5], [9], [22], [23]	[9]	Unknown	[6], [21]
	External Auditing Process	[7], [8], [11], [12], [13], [14], [15], [16], [25]	[8], [10], [25], [26]	[7], [8], [13], [16], [25], [26]	[7], [8], [10], [17], [24]	[10], [20], [27]	[6], [18], [19], [24]

[1] Amani & Fadlalla, (2017) [2] Chen & Du (2009) [3] Li, Sun & Wu (2010) [4] Sun & Li (2008) [5] Koh & Low (2004) [6] Peji´c Bach, Krstic, Seljan & Turulja (2019) [7] Kirkos, Spathis & Manolopoulos (2008) [8] Zhou & Kapoor (2011) [9] Dutta, Dutta & Raahemii (2017) [10] Ravisankar, Ravi, Rao & Bose (2011) [11] Busta & Weinberg (1998) [12] Bhattacharya, Xu & Kumar (2011) [13] Lin, Chiu, Huang & Yen (2015) [14] Green & Choi (1997) [15] Krambia-Kapardis, Christodoulou & Agathocleous (2010) [16] Kirkos, Spathis & Manolopoulos (2007) [17] Chen, Liou, Chen & Wu (2019) [18] West & Bhattacharya (2016) [19] Hajek & Henriques (2017) [20] Alden, Bryan, Lessley & Tripathy (2012) [21] Holton (2009) [22] Awodele, Akinjobi & Akinsola (2017) [23] Lin, Chiu, Huang & Yen (2015) [24] Gupta and Gill (2012) [25] Ngai, Hu, Wong, Chen & Sun (2011) [26] Liou (2008) [27] Welch, Reeves & Welch (1998)

Of the six DM methods, ANN and DT have attracted the greatest attention of researchers. According to studies, ANN and DT, which are data mining methods, are mostly preferred during audit activities. In 9 of the 27 studies examined, the ANN method is used in the external audit process. ANN is followed by DT with 6 studies. When the studies are examined in terms of the internal audit process, it has been determined that fewer DM methods are used than the external audit. In 5 of the 27 studies examined, the

DT method is used in the internal audit process. DT is followed by ANN with 4 studies. According to this result, it is seen that DM methods are used especially for external audit activities. However, the detection of these fraudulent activities is important and further investigation is essential. Accounting practitioners and auditors may consider investing in these methods to prevent fraud in their company and to respond to urgent regulatory demands and legal requirements required by the Sarbanes-Oxley Act (Lin et al. 2015).

4. Conclusion

A large amount of data emerges when companies carry out their activities, both during the period and when reporting at the end of the period. In order to detect fraudulent transactions or to make fraudulent transactions visible in the reporting, it becomes obligatory for the auditors to benefit from some statistical methods. Data mining is applied in many areas, including auditing activities. Data auditing tools can be used within data mining. Data mining methods demonstrated sufficient ability to model, classify and detect fraud. When data mining methods are considered holistically, there is a high expectation of assurance and compliance. In terms of audit activities, data mining methods especially focus on classification, estimation. Accurate fraud estimation is aimed in audit activity with data mining methods. Audit activities can be carried out more quickly, accurately, and cost-effectively with audit-based data mining tools.

DM practices reported in accounting contribute to audit activities in many ways, including well-founded technical justifications, incremental modeling approaches, unbiased model testing, and modeling of real-world problems (Amani and Fadlalla, 2017). DM methods provide benefits in many aspects during the auditing of financial statements. First, the increased emphasis on fraud detection in audits by regulators and standard setters provides motivation to identify and use tools to increase auditor efficiency. Second, the increasing use of DM tools as a forensic tool in accounting firms means that there is a growing population of people in firms with DM experience and a general awareness of DM. Third, due to the development of DM methods, easier to use and reliable tools have emerged (Gray and Debreceny, 2014).

In this study, data mining methods used to detect fraudulent transactions and to carry out audit activities more effectively are examined. In the study, the most frequently used data mining method was examined in detail. It was determined that some methods were used extensively in the auditing process. In the literature review, it was seen that artificial neural networks, logistic regression, decision trees, support vector machines, genetic algorithms, and text mining are used intensively in auditing activities. While the detection of fraud and fraudulent financial reports was examined within the scope of internal audit and external audit activities, it was determined that artificial neural networks and decision trees methods were mostly used. Ata and Seyrek (2009) similarly identified decision trees and artificial neural networks as two data mining techniques that are frequently used in the detection of financial statement fraud.

References

Ağdeniz, Ş. and Yıldız, B. (2018). Muhasebede analiz yöntemi olarak metin madenciliği. *Muhasebe Bilim Dünyası Dergisi*, *20*(2), 286-315.

AICPA. (1999). Top 10 technologies – plus 5 for tomorrow. *Journal of Accountancy*, *187*(5), 16-17.

Albizri, A., Appelbaum, D. and Rizzotto, N. (2019). Evaluation of financial statements fraud detection research: A multi-disciplinary analysis. *International Journal of Disclosure and Governance*, *16*(4), 206-241.

Alden, M., Bryan, D., Lessley, B. and Tripathy, A. (2012). Detection of financial statement fraud using evolutionary algorithms. *Journal of Emerging Technologies in Accounting*, *9*(1), 71-94.

Amani, F. and Fadlalla, A. (2017). Data mining applications in accounting: A review of the literature and organizing framework. *International Journal of Accounting Information Systems*, *24*, 32-58.

An, A. (2009). Classification methods. In: J. Wang, (Ed.), *Encyclopedia of Data Warehousing and Mining* (196-201). Hershey: IGI Global.

Ata, A. and Seyrek, İ. (2009). The use of data mining techniques in detecting fraudulent financial statements: An application on manufacturing firms. *Suleyman Demirel University Journal of Faculty of Economics & Administrative Sciences*, *14*(2), 157-170.

Awodele, O., Akinjobi, J. and Akinsola, J.E.T. (2017). A framework for web based detection of journal entries frauds using data mining algorithm. *International Journal of Computer Trends and Technology*, *51*(1), 1-9.

Bauder, R., Khoshgoftaar, T. and Seliya, N. (2017). A survey on the state of healthcare upcoding fraud analysis and detection, *Health Services and Outcomes Research Methodology, 17*(1), 31-55.

Bhattacharya, S., Xu, D. and Kumar, K. (2011). An ANN-based auditor decision support system using Benford's law. *Decision Support Systems*, *50*(3), 576-584.

Bhattacharyya, S., Jha, S., Tharakunnel, K. and Westland, C. (2011). Data mining for credit card fraud: A comparative study. *Decision Support Systems*, *50*(3), 602-613.

Busta, B. and Weinberg, R. (1998). Using Benford’s law and neural networks as a review procedure. *Managerial Auditing Journal*, *13*(6), 356-366.

Camps-Valls G., Martínez-Ramón M., Rojo-Álvarez, J. L. (2009). Applications of kernel methods. In: J. Wang, (Ed.), *Encyclopedia of Data Warehousing and Mining* (51-57). Hershey: IGI Global.

Cerchiello, P. (2009). Data mining and the text categorization framework. In: J. Wang, (Ed.), *Encyclopedia of Data Warehousing and Mining* (394-399). Hershey: IGI Global.

Chen, W. S. and Du, Y. K. (2009). Using neural networks and data mining techniques for the financial distress prediction model. *Expert Systems with Applications*, *36*(2), 4075-4086.

Chen, Y. J., Liou, W. C., Chen, Y. M. and Wu, J. H. (2019). Fraud detection for financial statements of business groups. *International Journal of Accounting Information Systems*, *32*, 1-23.

Chrysostomou, K., Lee, M., Chen, S. and Liu, X. (2009). Wrapper feature selection. In: J. Wang, (Ed.), *Encyclopedia of Data Warehousing and Mining* (2103-2108). Hershey: IGI Global.

Dalla Valle, L. (2009). Data mining for internationalization. In: J. Wang, (Ed.), *Encyclopedia of Data Warehousing and Mining* (424-430). Hershey: IGI Global.

Debreceny, R. and Gray, G. (2011). Data mining of electronic mail and auditing: a research agenda. *Journal of Information Systems, 25*(2), 195-226.

Dominik, A., Walczak, Z. and Wojciechowski, J. (2009). In: J. Wang, (Ed.), *Encyclopedia of Data Warehousing and Mining* (202-207). Hershey: IGI Global.

Dutta, I., Dutta, S. and Raahemii, B. (2017). Detecting financial restatements using data mining techniques. *Expert Systems with Applications*, *90*, 374-393.

Feroz, E. H., Kwon, T. M., Pastena, V. and Park, K. (2000). The efficacy of red flags in predicting the sec's targets: An artificial neural networks approach. *Intelligent Systems in Accounting, Finance & Management*, *9*(3), 145-157.

Gaganis, C. (2009). Classification techniques for the identification of falsified financial statements: a comparative analysis. *Intelligent Systems in Accounting, Finance & Management: International Journal*, *16*(3), 207-229.

Gepp, A., Linnenluecke, M., O'Neill, T. and Smith, T. (2018). Big data techniques in auditing research and practice: Current trends and future opportunities. *Journal of Accounting Literature*, *40*, 102-115.

Goel, S. and Gangolly, J. (2012). Beyond the numbers: Mining the annual reports for hidden cues indicative of financial statement fraud. *Intelligent Systems in Accounting, Finance and Management, 19*(2), 75-89.

Goldberg, D. E. (1989). *Genetic algorithms in search, optimization, and machine learning*. Reading, USA: Addison-Wesley.

González, P. C. and Velásquez, J. (2013). Characterization and detection of taxpayers with false invoices using data mining techniques. *Expert Systems with Applications*, *40*(5), 1427-1436.

Gray, G. and Debreceny, R. (2014). A taxonomy to guide research on the application of data mining to fraud detection in financial statement audits. *International Journal of Accounting Information Systems*, *15*(4), 357-380.

Green, B. P. and Choi, J. H. (1997). Assessing the risk management fraud through neural network technology. *Auditing A Journal of Practice & Theory*, *16*(1), 14-28.

Gupta, R. and Gill, N. S. (2012). Financial statement fraud detection using text mining. *International Journal of Advanced Computer Science and Applications*, *3*(12), 189-191.

Hajek, P. and Henriques, R. (2017). Mining corporate annual reports for intelligent detection of financial statement fraud-A comparative study of machine learning methods. *Knowledge-Based Systems*, *128*, 139-152.

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

Hassani, H., Huang, X., Silva, E. and Ghodsi, M. (2016). A review of data mining applications in crime. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *9*(3), 139-154.

Holton, C. (2009). Identifying disgruntled employee systems fraud risk through text mining: A simple solution for a multi-billion dollar problem. *Decision Support Systems*, *46*(4), 853-864.

Hsu, W. (2009). Evolutionary computation and genetic algorithms. In: J. Wang, (Ed.), *Encyclopedia of Data Warehousing and Mining* (817-821). Hershey: IGI Global.

Jans, M., Lybaert, N. and Vanhoof, K. (2010). Internal fraud risk reduction: Results of a data mining case study. *International Journal of Accounting Information Systems*, *11*(1), 17-41.

Jans, M., Van Der W., Jan M., Lybaert, N. and Vanhoof, K. (2011). A business process mining application for internal transaction fraud mitigation. *Expert Systems with Applications*, *38*(10), 13351-13359.

Keyvanpour, M. R., Javideh, M. and Ebrahimi, M. R. (2011). Detecting and investigating crime by means of data mining: A general crime matching framework. *Procedia Computer Science*, *3*, 872-880.

Kim, Y. and Vasarhelyi, M. (2012). A model to detect potentially fraudulent/abnormal wires of an insurance company: An unsupervised rule-based approach. *Journal of Emerging Technologies in Accounting*, *9*(1), 95-110.

Kirkos, E., Spathis, C. and Manolopoulos, Y. (2007). Data mining techniques for the detection of fraudulent financial statements. *Expert systems with applications*, *32*(4), 995-1003.

Kirkos, E., Spathis, C. and Manolopoulos, Y. (2008). Support vector machines, decision trees and neural networks for auditor selection. *Journal of Computational Methods in Sciences and Engineering*, *8*(3), 213-224.

Koh, H. C. and Low, C. K. (2004). Going concern prediction using data mining techniques. *Managerial Auditing Journal*, *19*(3), 462-476.

Krambia-Kapardis, M., Christodoulou, C. and Agathocleous, M. (2010). Neural networks: The panacea in fraud detection? *Managerial Auditing Journal*, *25*(7), 659-678.

Kurban, O. C., Niyaz, Ö. and Yıldırım, T. (2016). Neural network based wrist vein identification using ordinary camera. In *2016 International Symposium on Innovations in Intelligent Systems and Applications (INISTA)*, Sinaia, Romania, 1-4.

Lewis, R. and Ras, Z. (2009). Facial recognition. In: J. Wang, (Ed.), *Encyclopedia of Data Warehousing and Mining* (857-862). Hershey: IGI Global.

Li, H., Sun, J. and Wu, J. (2010). Predicting business failure using classification and regression tree: An empirical comparison with popular classical statistical methods and top classification mining methods. *Expert Systems with Applications*, *37*(8), 5895-5904.

Lin, C. C., Chiu, A. A., Huang, S. Y. and Yen, D. (2015). Detecting the financial statement fraud: The analysis of the differences between data mining techniques and experts' judgments. *Knowledge-Based Systems*, *89*, 459-470.

Lin, S. C. and Lehto, M. (2009). A bayesian based machine learning application to task analysis. In: J. Wang, (Ed.), *Encyclopedia of Data Warehousing and Mining* (133-139). Hershey: IGI Global.

Liou, F. M. (2008). Fraudulent financial reporting detection and business failure prediction models: A comparison. *Managerial Auditing Journal*, *23*(7), 650-662.

Ngai, W. T. E., Hu, Y., Wong, Y. H., Chen, Y. and Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*, *50*(3), 559-569.

Peji´c Bach, M., Krstic, Ž., Seljan, S. and Turulja, L. (2019). Text mining for big data analysis in financial sector: A literature review. *Sustainability*, *11*(5), 1-27.

Ravisankar, P., Ravi V., Rao, G. R. and Bose, I. (2011). Detection of financial statement fraud and feature selection using data mining techniques. *Decision Support Systems*, *50*(2), 491-500.

Rivero, D., Rabuñal, J., Dorado, J. and Pazos, A. (2009). Evolutionary development of ANNs for data mining. In: J. Wang, (Ed.), *Encyclopedia of Data Warehousing and Mining* (829-835). Hershey: IGI Global.

Saxena, A., Kothari, M. and Pandey, N. (2009). Evolutionary approach to dimensionality reduction. In: J. Wang, (Ed.), *Encyclopedia of Data Warehousing and Mining* (810-816). Hershey: IGI Global.

Siciliano, R. and Conversano, C. (2009). Decision tree induction. In: J. Wang, (Ed.), *Encyclopedia of Data Warehousing and Mining* (624-630). Hershey: IGI Global.

Sun, J. and Li, H. (2008). Data mining method for listed companies' financial distress prediction. *Knowledge-Based Systems*, *21*(1), 1-5.

Welch, O., Reeves, T. and Welch, S. (1998). Using a genetic algorithm-based classifier system for modeling auditor decision behavior in a fraud setting. *International Journal of Intelligent Systems in Accounting, Finance and Management*, *7*(3), 173-186.

West, J. and Bhattacharya, M. (2016). Intelligent financial fraud detection: A comprehensive review. *Computers & Security*, *57*, 47-66.

Xiao, M., Xiaoli, H. and Gaojin, L. (2010). Research on application of data mining technology in financial decision support system. In *2010 3rd International Conference on Information Management, Innovation Management and Industrial Engineering*, *4*, 381-384.

Yigitbasioglu, O. and Velcu, O. (2012). A review of dashboards in performance management: Implications for design and research. *International Journal of Accounting Information Systems*, *13*(1), 41-59.

Yildiz, B. and Yezegel, A. (2010). Fundamental analysis with artificial neural network. *The International Journal of Business and Finance Research*, *4*(1), 149-158.

Yom-Tov, E. (2003). An introduction to pattern classification decision tree induction. In: O. Bousquet, U. Von Luxburg and G. Rätsch, (Ed.), *Advanced Lectures on Machine Learning* (1-20). Berlin, Heidelberg: Springer.

Yu, Z., Guang, Y. and Zi-qi, J. (2013). Violations detection of listed companies based on decision tree and K-nearest neighbor. In *2013 International Conference on Management Science and Engineering 20th Annual Conference Proceedings*, 1671-1676.

Zhou, W. and Kapoor, G. (2011). Detecting evolutionary financial statement fraud. *Decision Support Systems*, *50*(3), 570-575.

Zhu, D. (2009). Analytical competition for managing customer relations. In: J. Wang, (Ed.), *Encyclopedia of Data Warehousing and Mining* (25-30). Hershey: IGI Global.