sciendo

The KDD Process in Big Data Analytics: A Theoretical Approach to Taxpayer Non-Compliance Analysis

Arnela Kaknjo
University of Sarajevo – School of Economics and Business
arnela.kaknjo@gmail.com

Lejla Turulja
University of Sarajevo – School of Economics and Business
lejla.turulja@efsa.unsa.ba

Abstract

In the modern business environment, big data analytics and data mining techniques are increasingly recognized as tools for improving fiscal discipline and more efficient management of public revenues. This paper explores the possibility of applying the knowledge discovery process from databases to detect patterns of financial behavior that may indicate tax non-compliance. A quantitative approach based on the analysis of secondary data from ten joint-stock companies from the Federation of Bosnia and Herzegovina, for which financial statements and tax debt data are available, was used.

Paper type: Research article

Received: June 15, 2025
Accepted: June 30, 2025

Citation: Kaknjo, A., & Turulja, L. (2025). The KDD Process in Big Data Analytics: A Theoretical Approach to Taxpayer Non-Compliance Analysis. *Journal of Forensic Accounting Profession, 5*(1), pp. 16 - 42

DOI: https://doi.org/10.2478/jfap-2025-0002

The relationship between key financial indicators (EPS, financial stability ratio, total asset turnover ratio and debt ratio) and the amount of tax debt was examined using descriptive statistics and regression analysis. The results show that lower profitability and poorer financial stability significantly correlate with higher tax debt, while high operational efficiency and debt have a more complex and statistically marginal impact. The findings confirm the possibility of using publicly available financial data for early identification of risky taxpayers, which opens up space for further development of predictive models in the domain of tax analytics.

Keywords: *tax non-compliance, tax debt, data mining, financial stability, earnings per share, indebtedness, asset turnover,regression analysis, KDD process*

1. Introduction

The emergence of the Internet, networks, mobile technologies, as well as the mass digitization of business processes, has led to an exponential growth in the amount of raw data. In order to efficiently manage such massive and rapidly growing information flows, innovative programming models and systems have been developed that enable simplified processing of large data sets, expressed in terabytes or even petabytes. Also, open-source technical solutions have been developed to provide cost-effective data storage and processing, and working with big data is now a routine operation practically every day, simple and widely available to many organizations. The term “big data” itself refers to massive data sets that are extensive, diverse and complex, which makes their storage, analysis and visualization for further processing and decision-making difficult (Sagiroglu and Sinanc, 2013). In this regard, it is data that is too large and/or complex to be processed efficiently and/or effectively with traditional theories, technologies and tools (Cao, 2017).

However, big data is not defined solely by volume, but also by other dimensions, most commonly variety and velocity. These three “V” characteristics (volume, variety and velocity) have emerged as a common framework for describing big data (Gandomi and Haider, 2015). Volume refers to the size of data, depending on factors such as time and data type, and is expressed in multiple terabytes and petabytes (Gandomi and Haider, 2015). However, with large volumes of

data comes the problem of storing and managing them (Brady, 2019). Variety represents the structural heterogeneity in the data set (Gandomi and Haider, 2015), but increasing variety also makes it difficult to translate data between different forms and formats (Brady, 2019). Velocity refers to the rate at which data is generated and the speed at which it needs to be analyzed and acted upon (Gandomi and Haider, 2015). The accelerated pace of information generation imposes the need for its processing in real time and calls into question the credibility of the data, which is further complicated by data management and makes the analysis process even more complex (Brady, 2019). The exploration of big data with the aim of identifying hidden patterns and their mutual correlations is called big data analytics (Sagiroglu and Sinanc, 2013). Data analytics, or analytical methods, are divided according to the type of data and their function into (Gandomi and Haider, 2015): text analytics, video analytics, social network analysis, and predictive analytics. Data analytics uses data science as the foundation for its methods and is, at the same time, an integral part of the broader process that data science encompasses. Data science does not only include data analysis, but is an interdisciplinary field that brings together statistics, informatics, computer science, communication, management and sociology with the aim of transforming data into useful insights and informed decisions (Cao, 2017). In this context, Brady (2019) positions data science at the intersection of programming skills, mathematical-statistical knowledge and expertise in a specific field of research. Specifically, Igual and Seguí (2024) define data science as a methodology that can be used to derive actionable insights from data. Through the process of discovering, formulating, and testing hypotheses, data science enables the extraction of actionable knowledge directly from data (Brady, 2019).

The fundamental principles underlying data science provide a systematic approach to extracting information and knowledge from data. Using the principles, processes, and techniques of analytics, data science also provides understanding of various phenomena (Provost and Fawcett, 2013), a deeper understanding of human behavior, and support for decision-making (Igual and Seguí, 2024). In this regard, discoveries, predictions, recommendations, insights for decision-making, concepts of thinking, models, paradigms, tools, or systems are all products that result from, or are enabled and driven by, data. However, the end product, or the most valuable form of results from data, is reflected in the creation of knowledge, intelligence, wisdom, and decision-making (Cao, 2017). The main goal of data science is to form conclusions based on data, which serve as a basis for decision-making. Without data, beliefs are uninformed, and decisions rely on

intuition or established practices at best. Representing complex systems through the richness of data opens up the possibility of applying scientific methods and principles in the process of deriving knowledge from data (Igual and Seguí, 2024).

The success of data science does not depend solely on data mining algorithms, but also on the ability of analysts to view business challenges from the perspective of data (Provost and Fawcett, 2013). Companies that have recognized the importance, that is, the value and potential of data science and analytics, have strategically differentiated themselves and used data to create new business opportunities and increase productivity (Cao, 2017). Given that it is difficult to precisely define the boundaries of data science, it is important to understand its relationship with other related concepts and identify the key principles on which it is based (Provost and Fawcett, 2013). In this context, Igual and Seguí (2024) point out that its application is not limited to business, but extends to social sciences, management, as well as other non-traditional fields. Finaly, data science can be considered as a discipline that allows the discovery of new and significant relationships, patterns and trends when analyzing large amounts of data. Accordingly, data science techniques such as clustering, classification, predictive analytics, mining and others that aim to discover patterns, profiles and trends through data analysis, strive to automatically find knowledge contained in information stored in large databases (Fernández et al., 2018).

However, despite the widespread application of big data analytics in numerous sectors, the application of KDD processes and data mining techniques in the context of tax (non)compliance remains relatively under-researched, especially when it comes to integrating financial indicators with actual tax debt records. Most existing research focuses on general risk and fraud analytics, but does not detail specific indicators that could anticipate tax non-compliance through quantitative models. This research gap is reflected in the limited number of empirical works that link corporate financial performance to the occurrence of tax debts through the application of data mining and regression analyses, especially in regional and local contexts such as the case of the Federation of Bosnia and Herzegovina. Given the above, the goal of this research is twofold. First, the goal is to systematize and demonstrate how the KDD process functions within the framework of big data analytics, with a special focus on mining techniques that can be used to detect organizational non-compliance. Second, the goal of the research is to demonstrate that certain financial indicators such as earnings per share, financial stability ratio, leverage and asset turnover

can be statistically linked to the occurrence of tax debt, thus enabling predictive detection of potential tax non-compliance using simple analytical models.

To examine the research questions, this paper uses a quantitative approach based on secondary data analysis. The focus is on joint-stock companies in the Federation of BiH for which financial statements and tax debt data are available. Using descriptive statistics and regression analysis, the relationship between key financial indicators and the amount of tax debt is investigated. This approach allows for the identification of potential patterns of tax non-compliance and testing the hypothesis that certain characteristics of financial operations may indicate an increased risk of tax non-compliance. By linking this empirical approach to the process of knowledge discovery in databases, the study shows how data mining techniques can be applied in a real regulatory context to extract useful insights from structured financial data. In this way, it bridges the gap between the theoretical foundations of big data analytics and their practical application in tax non-compliance detection, offering a simplified but illustrative implementation of the KDD methodology.

2. Literature review

2.1. Knowledge discovery from large databases

In today's era of widespread digitalization, huge amounts of data are collected daily in almost every segment of human activity, be it business, science, social networks, trade, production, etc. However, the mere existence of data does not mean the presence of knowledge. Therefore, in recent years, there has been a need to implement processes in business that will respond to the accelerated digital transformation and growing volumes of data and convert data into valuable knowledge needed to make informed decisions and optimize business. Knowledge Discovery in Databases (KDD) is the non-trivial extraction of previously unknown and potentially useful information from a database (Frawley et al., 1992). It is a multi-phase process that begins with raw data that, through a series of transformations and cleaning, is converted into information that can be used to support decision-making and improve business process efficiency (Fayyad et al., 1996).

In addition, enriching the KDD process with domain-specific expertise enables the discovery of relevant and specific knowledge in real-world business environments, thereby improving the usefulness and relevance of data, as well as providing deeper and more meaningful insights into the patterns discovered (Cao and Zhang, 2007). Unlike static analyses, the KDD process focuses

on discovering knowledge from data streams, whereby continuous collection and analysis of real-time data gives greater weight and significance to decision-making and creates more relevant and efficient models in dynamic environments, unlike static analyses based on stored or outdated data (Gama et al., 2008).

KDD uses statistics, databases, and artificial intelligence to develop tools that provide understandable insights from large data sets, and the knowledge that is created can have practical applications in various fields. It should be noted that KDD is not solely technical in nature; it also requires a combination of technology and human perception in order to better understand the needs of users and their understanding of insights from large databases (Pazzani, 2000). Therefore, KDD represents a comprehensive process for developing business knowledge and decision-making that cannot fully realize its potential without collaboration between stakeholders and continuous knowledge sharing throughout the process (Wang and Wang, 2008).

In order for the patterns discovered through KDD to be valid and relevant beyond the initial data set, they need to be applicable to new data with a certain degree of reliability. Accordingly, KDD aims to discover non-trivial patterns, i.e. patterns that cannot be easily calculated or predicted in advance. Understandability and usefulness of patterns are criteria that ensure that the discovered patterns are not only statistically significant, but also practically useful, and therefore applicable in real situations (Gullo, 2015).

The sequence of tasks are performed in the KDD process represents the phases of its life cycle, which according to Rahman et al. (2014) includes the following key steps:

- Understanding the business context in which the company's goals and requirements are defined;
- Understanding the data, i.e. exploring its characteristics;
- Preparing the data, collecting it, cleaning, transforming and formatting it;
- Modeling through the application of analytical methods and data mining algorithms;
- Evaluating the models and checking their applicability;
- Applying the discovered knowledge and using the results in real situations.

A clear definition of the business context in the first phase of the KDD process determines the specific goals of data analysis and the company's requirements that are sought to be achieved (Rahman et al., 2014). Therefore, understanding the data itself depends on the ability to explore the characteristics of the data, as well as prior knowledge about the business domain (Zemmouri et al., 2012). A system based on such a framework improves the process of knowledge discovery throughout its entire life cycle (Cao and Zhang, 2007). Data preparation, selection, cleaning, transformation and formatting are steps that aim to present relevant data in a form suitable for the application of analytical methods (Rahman et al., 2014). This is data preprocessing (Fayyad et al., 1996) in which data "cleaning" is performed (Gullo, 2015), and thus the quality of raw data is improved (Fan et al., 2021).

Data mining is a fundamental step in the knowledge discovery process in which, by applying intelligent methods and algorithms for pattern extraction, data analytics is carried out. Data mining can be applied to various types of data, and the most basic forms used are data from databases, data warehouses and transactional data (Han et al., 2012). The patterns and models generated by data mining algorithms are subjected to detailed analysis in the evaluation and interpretation phase, with the aim of assessing their validity. At the same time, the user also evaluates the practical value of the discovered knowledge and its applicability in a specific context and real-world situations (Ristoski and Paulheim, 2016). According to Rahman et al. (2014), this step is iterative, meaning that the results of the analysis are often used as feedback for improving the model and further optimizing the process.

2.2. Challenges of the knowledge discovery process from large databases

The KDD process is frequently confronted with challenges that hinder its effective implementation, such as data quality problems, complexity of preprocessing, integration of heterogeneous sources, existence of scalable and efficient algorithms, etc. In addition, real-world data and databases are dynamic and complex, and are often incomplete, incorrect or contain erroneous data, which emphasizes the need for more advanced algorithms, and therefore additional steps in data preprocessing and filtering (Fayyad et al., 1996). Therefore, it is wrong to expect that the system should work correctly only when there are no errors in the entered data, but it is necessary to provide tools for error recognition and correction (Phalgune et al., 2005). Namely, data mining processes depend on the accuracy and quality of input data, and the results of applying

these methods to poor-quality or inconsistent data cannot be considered reliable (Arora et al., 2009).

In addition, the increasing volumes of data requiring analysis, i.e. the huge influx of data that do not fit into traditional concepts of data structure, pose a problem for knowledge discovery and data mining techniques (Usai et al., 2018; EMC, 2012). Also, the very nature of the analyzed data, such as its distribution, heterogeneity and incompleteness, adds further complexity to the KDD process (Zemmouri et al., 2012). The heterogeneity of big data arises as a result of unstructured data that limits the potential opportunities for achieving accuracy in the KDD process (Lomotey and Deters, 2014). Therefore, heterogeneous data coming from different sources, i.e. data of different content, context and structure, is challenging to meaningfully unify into a single research database (Kubick, 2012), (Kamm et al., 2021). In this context, there is also the challenge of scalability of algorithms that are designed for smaller data sets, which is why they face difficulties when using them on large data sets (Piatetsky-Shapiro, 1990). According to Gama et al. (2008), the challenge faced by the KDD process is also reflected in the selection of appropriate data mining techniques, bearing in mind that the process phases themselves bring a series of choices that determine the final outcome of the project, and the decision on the method can significantly affect both the quality and the interpretation of the discovered patterns. Finally, Wang and Wang, (2008) emphasize the importance of understanding the specific circumstances in which the tools and techniques of the KDD process are applied, and for the correct application of the techniques, and therefore for the correct interpretation of the results and their application in the environment, understanding the context is crucial, as well as the active involvement of end users, which is achieved through their cooperation and sharing of knowledge with analysts.

2.3. The role of big data analytics in the KDD process and mining techniques

According to Cao (2017), analysis, analytics and advanced data analytics are different concepts within data science, since each of them has a specific role and depth of approach in data processing. Namely, data analysis uses traditional theories, technologies and tools to obtain useful information and practical insights. On the other hand, data analytics represents theories, technologies, tools and processes that enable in-depth discovery of applicable insights from data, while advanced analytics enables discovery and in-depth understanding of applicable insights from big data. A key role within the KDD process is played by data analytics, which improves the study and extraction of

information from data. Data analytics consists of descriptive, predictive and prescriptive analytics. Descriptive analytics is a type of data analytics that uses statistics to describe data in order to gain information. Predictive analytics forecasts future events and identifies underlying drivers or contributing factors, while prescriptive analytics recommends optimal actions based on predictive insights to support decision-making (Cao, 2017).

When talking about big data, big data analytics relies on various data analytics methods to discover new knowledge such as clustering, classification, and association rules, many of which fall under the domain of data mining (Dedić and Stanier, 2017). Data mining is the fundamental process of discovering useful and interesting patterns, models, and other forms of knowledge from large data sets. This process converts vast and complex datasets into actionable knowledge. Data mining applies intelligent methods to extract patterns, create models, or to derive knowledge in various forms, depending on specific mining functions and use cases (Han et al., 2022). Also, data mining tasks can be classified into two categories: descriptive mining and predictive data mining. Descriptive mining describes the properties of a selected data set, while predictive mining performs induction on a data set to make predictions (Han et al., 2022). Data mining as a process of knowledge discovery has been explained through numerous studies covering methods such as classification, clustering, regression, anomaly detection, text mining, and deep learning.

2.4. Tax (non)compliance

The economics of tax compliance and tax law compliance can be approached from several perspectives, more precisely, it can be viewed as a problem of public finance, a legal problem of law enforcement, organizational structure, workforce or ethics, or a combination of all of the above (Andreoni et al., 1998). Tax compliance, sometimes referred to as adherence to tax law, is a neutral term that defines the willingness of taxpayers to pay taxes (Kirchler and Wahl, 2010). According to the generally accepted opinion, the simplest form of tax compliance is set in terms of the degree to which taxpayers respect and comply with tax laws (Roth and Scholz, 1989). In Taxpayer Compliance, Volume 1: An Agenda for Research, Roth and Scholz (1989) state that compliance implies that the taxpayer files all required tax returns on time and that the returns accurately reflect the tax liability in accordance with the tax law, regulations, and other laws and regulations in effect at the time of filing. Otherwise, non-compliance occurs. Non-compliance refers to the degree of

"tax gap," which is the difference between the income actually collected and the amount that would have been collected if there was 100 percent compliance (James and Alley, 2002).

One of the first and perhaps most famous economic models of tax compliance was presented by Michael G. Allingham and Agnar Sandmo, according to which the taxpayer chooses whether to report the income he has earned accurately or to choose a strategy in which he reports less than he has actually earned, while taking the risk of being detected by tax authorities. If his return is not subject to investigation, his strategy is all the more successful. However, if he is investigated, an audit may result in a penalty. In this way, the taxpayer chooses the amount to declare, in order to maximize the expected utility from taking the risk of evasion (Allingham and Sandmo, 1972). Furthermore, compliance can be voluntary or forced by the competent authorities. Failures to fulfill tax obligations, whether intentional or not, result in either legal tax avoidance or illegal tax evasion, depending on the intent and nature of the action (Kirchler and Wahl, 2010). Sandmo and Allingham (1972) define tax evasion as an attempt by a taxpayer to free himself from the tax burden without any legal basis for such an exemption.

In addition to avoiding paying taxes on income, acquired wealth or any other income, according to (Myles, 2000) tax evasion also includes any illegal reduction of them through financial statements, i.e. tax returns. In this context, Zandi et al. (2019) explains that understating the profits made, any form of inaccurate or deliberately misleading tax reporting, illegally conceals the true income status of the taxpayer. Therefore, whether it is understating or overstating the tax liability, it is in any case a tax non-compliance (Roth and Scholz, 1989).

2.5. The potential for big data analytics to detect tax non-compliance

Modern administrations around the world face complex challenges in identifying and collecting public revenues from businesses that engage in sophisticated methods of tax avoidance. Given limited resources and traditional tax audit strategies, authorities are finding it increasingly difficult to respond to innovative forms of tax non-compliance and evasion. On the other hand, companies are increasingly recognizing the importance of data analytics as a strategic resource not only for improving business results, but also for managing regulatory risks, including the risk of tax fraud. Therefore, there is an evident need for more efficient and data-driven approaches to identifying risky behavior patterns, and data mining techniques offer one of the more advanced

solutions for detecting and preventing irregularities in tax returns and transactions. Wahab and Bakar (2021) point out that big data technology has become more accessible, simpler, and more financially acceptable with the use of data mining in tax compliance analysis. Namely, the legitimacy and intent behind a tax filing cannot be determined solely based on its appearance or the taxpayer's profile. Therefore, the most cost-effective option is to extract possible indicators of fraudulent declarations or declarations from available data using data mining algorithms (Gupta and Nagadevara, 2007).

Data mining enables the extraction of possible indications of fraudulent tax filings and the recognition of hidden signs of fraud through algorithms that humans or basic tools would not notice. The predictive features of their models find taxpayers who underestimate or avoid tax filing and, due to the reduction in the need to review all returns, the result is reduced operational costs and improved detection efficiency in detecting irregularities (Gupta and Nagadevara, 2007). In this regard, Wahab and Bakar (2021) propose machine learning algorithms to develop classification models for identifying compliant and non-compliant taxpayers. Namely, predictive analytical models find key characteristics, or attributes, that contribute most to classifying entities into categories of compliant and non-compliant taxpayers.

In addition, Wu et al. (2012) developed a selection framework to filter out potentially non-compliant VAT returns that could be subject to further audit. The results show that the associative rule mining of data improves the detection of tax evasion and can be effectively applied to reduce or minimize losses caused by tax evasion. Also, the ability to apply big data analytics allows companies to strategically use data to strengthen organizational capabilities and achieve competitive advantages, and the targeted use of big data analytics is crucial for organizations to ensure long-term sustained competitive advantage (Mikalef et al., 2018). Empirical research confirms that companies that have big data analytics capabilities achieve better performance in the market and at the operational level compared to their competitors, more precisely, the ability to apply big data analytics leads to superior business results for companies (Gupta and George, 2016).

However, developing the ability to apply big data analytics involves the effective alignment of tangible, intangible and human resources, which includes their restructuring, connection and strategic use with the aim of strengthening organizational capacities at the enterprise level (Mikalef

et al., 2018). Gupta and George (2016) also emphasize that the achievement of superior company results is enabled by a specific combination of tangible resources such as data, technology, investments, then human skills such as managerial and technical expertise and intangible resources such as data-driven culture and organizational learning. Accordingly, the ability to develop infrastructure in terms of IT, managerial capabilities to manage IT resources in accordance with business needs and staff capabilities, that is, the professional skills and knowledge of employees to perform tasks related to data analytics, are three components that together enable companies to use big data analytics (Wamba et al., 2017).

In light of the above, it becomes apparent that the ability to effectively apply big data analytics is not only a source of competitive advantage for companies, but also a valuable tool for tax administrations seeking to detect non-compliant behavior. The literature emphasizes that combining financial, organizational, and technological capabilities is crucial for successfully extracting actionable insights from complex and voluminous data sets. This perspective supports the idea that big data analytics, especially through the use of machine learning and data mining techniques, can be applied beyond commercial use as a tool for public interest, such as identifying indicators of tax non-compliance.

2.6. Theoretical rationale for linking financial indicators and tax non-compliance

Previous research on tax compliance has mainly focused on behavioral, institutional, and legal aspects (Boateng et al., 2022), while less attention has been paid to the potential of quantitative financial indicators to serve as early indicators of tax non-compliance. Theoretical frameworks such as liquidity disruption theory and pecking order theory suggest that financial pressures and structural problems in business can affect a firm's tax compliance behavior (Boateng et al., 2022; Seidu et al., 2023). For example, lower profitability may indicate a reduced ability of a firm to meet its tax obligations. Similarly, capital structure imbalances and high leverage may be signs of reduced financial stability, which increases the risk of late or evasion of tax payments. Although asset turnover indicates operational efficiency, in the context of tax liabilities it can have a neutral or even negative impact, especially when not accompanied by responsible liability management or in conditions of financial pressure (Boateng et al., 2022). Financial constraints can have a dual effect. On the one hand, they threaten the daily operations of the company, and on the other hand, they negatively affect public revenues through reduced tax revenue (Agyei & Yankey, 2019; Seidu

et al., 2023). When companies face limited financial resources, they often resort to various management strategies to maximize operational capacity, including various forms of tax planning (Seidu et al., 2023). Financial constraints occur when a company becomes increasingly expensive and difficult to access external sources of financing, or when it cannot secure the necessary levels of external funds for its operations (Seidu et al., 2023).

These conceptual insights form the basis for the following hypothesis:

H: There is a statistically significant relationship between specific financial metrics and the presence of tax debts.

This hypothesis is based on the basic principles of the KDD process, whereby financial data is systematically processed and analyzed using data mining techniques to uncover significant patterns, especially indicators of potential tax non-compliance.

3. Empirical research

A quantitative research design was adopted in this study, relying on the analysis of secondary data. The primary objective was to determine whether specific financial characteristics of companies can be statistically associated with the presence of tax debt, and whether such patterns can be identified through data mining techniques to predict the likelihood of tax non-compliance, either through avoidance or failure to fulfill tax obligations.

3.1. Data source and sample selection

The core dataset was derived from the public registry published by the Tax Administration of the Federation of Bosnia and Herzegovina on January 22, 2024, listing all taxpayers with debts exceeding 50,000 BAM as of December 31, 2023. This tabular report included the following variables: taxpayer number, company name, identification number, registered office, cantonal tax office, debt amount (principal and interest), and the outstanding balance on the given date. The dataset comprised 4,989 taxpayers with a total outstanding debt of 2,703,486,282.80 BAM, of which 2,073,039,499.64 BAM represented the principal and 630,446,783.16 BAM the interest. Among these entities, 4,927 unique identification numbers were confirmed, while 62 taxpayers could not be reliably matched due to missing or unverifiable identifiers.

The largest number of indebted taxpayers were located in Sarajevo (1,287), followed by Tuzla (466), Mostar (337), Bihać (246), and other municipalities. The types of organizations varied, but most of the debt was attributed to limited liability companies (3,150), sole proprietorships (852), and joint-stock companies (191). The remaining debts were linked to public institutions, associations, cooperatives, and inactive firms (582), for which no active registration or legal status was found. For the purpose of this research, joint-stock companies were selected as the focus group due to their legal obligation to publicly disclose financial statements. Based on the availability of such statements, a sample of 10 joint-stock companies was identified, for which complete 2023 financial reports (balance sheet and income statement) were publicly accessible via the official website of the Sarajevo Stock Exchange (SASE). Financial data for other joint-stock companies were either not publicly available or had not been published at the time of data collection.

3.2. Data matching and preprocessing

To merge tax debt data with financial statement information, a unique company identifier was used as a matching key. The tax identification number (JIB), available in both datasets, was employed to ensure precise alignment across data sources. Records lacking valid identifiers were removed during the data cleaning phase. Further preprocessing included the elimination of duplicates, normalization of company names (adjusting for case sensitivity, spacing, and special characters), and temporal validation to ensure that financial statements corresponded to the same reporting period as the tax debt data. Given that tax debt values reflected the situation as of December 31, 2023, only financial reports covering the period from January 1 to December 31, 2023 were included.

3.3. Financial indicators

As part of data preparation, the following financial indicators were calculated to capture various operational dimension: profitability, financial stability, leverage, and activity, which have been theoretically and empirically linked to tax debt levels and the likelihood of non-compliance:

- Earnings Per Share (EPS): net income / number of ordinary shares
- Financial Stability Ratio: non-current assets / (equity + long-term liabilities)

- Asset Turnover Ratio: total revenue / total assets
- Debt Ratio: total liabilities / total assets

The EPS (Earnings Per Share) shows how much profit is attributable to each ordinary share. A higher or growing EPS is often interpreted as a sign of strong financial performance, while a low or declining EPS may indicate financial difficulties.

The Financial Stability Ratio reflects the company's ability to maintain a healthy balance between debt and equity, calculated by dividing non-current assets by the sum of equity and long-term liabilities. This ratio is commonly used as an indicator of liquidity and solvency, affecting investor confidence and access to capital markets (Škaro, 2024).

The Asset Turnover Ratio measures how effectively a company uses its assets to generate revenue. Specifically, how many units of revenue are produced for each unit of asset value. A rising trend in this ratio is generally considered favorable, as prior studies suggest a negative correlation between asset turnover and the probability of insolvency (Šarlija et al., 2009).

Finally, the Debt Ratio shows the extent to which company assets are financed by external debt. It is calculated as the ratio of total liabilities to total assets. A higher debt ratio indicates greater reliance on borrowed capital, and therefore a higher financial risk, especially in terms of the firm's ability to meet future obligations (Barić, 2022).

4. Data analytics

4.1. Variable standardization and data overview

To analyze the relative importance of predictors while maintaining economic interpretability, the dependent variable, tax debt, was kept in its actual monetary units (BAM). This approach ensures a more meaningful interpretation of the regression results. The EPS (Earnings Per Share) variable already possesses its own natural economic unit (BAM per share), and therefore, no additional normalization was required. On the other hand, to enable meaningful comparative analysis across the independent variables, i.e., financial indicators, and to express all predictors in the same standardized units, Z-score normalization was applied. In this process, each value was expressed as a deviation from the arithmetic mean in units of standard deviation. This approach

allowed for comparable measurement of different financial ratios across the sample, regardless of their original scale, thereby ensuring homogeneity of the data in terms of variability.

Table 1: Overview of variables used in the analysis

No.	Tax Debt (BAM)	EPS (BAM)	Financial Stability Ratio (Z)	Asset Turnover Ratio (Z)	Debt Ratio (Z)
1	15,342,580.22	-58.3024	-2.4187	-1.1575	1.1358
2	11,496,434.31	0.7772	-1.1062	0.4414	0.3210
3	1,312,122.86	-1.5270	0.5659	0.4987	-0.5197
4	5,652,754.97	-34.8870	0.4605	0.4317	-0.5783
5	3,267,943.36	0.0207	0.5459	-0.7009	-0.7803
6	552,564.70	-0.0223	0.2859	-1.1697	-1.4692
7	1,329,103.06	-4.6347	0.8274	1.6830	0.1996
8	872,413.79	0.0112	0.4828	-1.0920	-0.9003
9	211,767.00	1.4056	0.3062	1.0705	1.4120
10	721,807.06	-3.4068	0.0503	-0.0050	1.1794

Source: Authors' calculation

4.2. Regression analysis – Model I: EPS as predictor

A simple linear regression analysis was first conducted to assess the relationship between EPS (independent variable) and tax debt (dependent variable). This approach allowed for a focused and clear test of the hypothesis regarding the direct association between company profitability and tax liabilities, without interference from other financial variables. Given the limited sample size of ten joint-stock companies, simple regression was deemed methodologically appropriate, as it minimizes the risk of overfitting and misinterpretation. Additionally, this analysis provides a foundation for more advanced and complex modeling in future research.

Regression model I:

$$TaxDebt_i = \beta 0 + \beta 1 * EPS_i + \varepsilon i$$

Where:

- *TaxDebti*: dependent variable, tax debt of company *i*
- *β0*: intercept, expected tax debt when EPS = 0
- *β1*: regression coefficient, indicating the change in tax debt for a one-unit change in EPS

- *EPSi*: independent variable, earnings per share for company *i*
- *εi*: error term, representing other factors not included in the model

Table 2: Regression results – Model I

Regression Statistics	Value
Multiple R	0.7170
R Square	0.5141
Adjusted R Square	0.4534
Standard Error	3,890.3149
Observations	10

Source: Authors' calculation

With 10 observations, the regression model yielded a multiple correlation coefficient of 0.717, indicating a moderate to strong positive linear relationship between EPS and tax debt. The R-squared value of 0.514 suggests that approximately 51.4% of the variance in tax debt can be explained by the EPS variable.

Table 3: ANOVA results – Model I

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	128.106.314,0213	128.106.314,0213	8,4645	0,0196
Residual	8	121.076.400,4436	15.134.550,0554		
Total	9	249.182.714,4649			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 80,0%*	*Upper 80,0%*
Intercept	2.190,0935	1.390,5453	1,5750	0,1539	-1.016,5097	5.396,6966	247,7586	4.132,4284
EPS	-187,5253	64,4554	-2,9094	0,0196	-336,1597	-38,8909	-277,5576	-97,4930

Source: Authors' calculation

The ANOVA test indicated that the model is statistically significant, with an F-value of 8.46 and a corresponding p-value of 0.0196, which is below the conventional significance threshold of 0.05. This implies that the relationship between EPS and tax debt is statistically meaningful.

Regarding the regression coefficients, the coefficient for EPS was negative (-187.53) and statistically significant (p = 0.0196), suggesting that an increase in EPS by one unit is associated with a decrease in tax debt by approximately 187.53 BAM, holding other factors constant. The

intercept was not statistically significant (p = 0.1539), indicating that its interpretation when EPS = 0 may not be meaningful.

4.3. Regression analysis – Model II: multiple financial indicators

Regression model II:

$$TaxDebt_i = \beta 0 + \beta 1 * FSr_i + \beta 2 * AT_i + \beta 3 * DR_i + \varepsilon i$$

Where:

- *TaxDebti*: tax debt of company i
- *FSri*: financial stability ratio
- *ATi*: asset turnover ratio
- *DRi*: debt ratio
- *β0*: intercept
- *β1, β2, β3*: regression coefficients
- *εi*: error term

Table 4: Regression results – Model II

Regression Statistics	Value
Multiple R	0.9569
R Square	0.9157
Adjusted R Square	0.8736
Standard Error	1,871.0933
Observations	10

Source: Authors' calculation

This table presents the key performance indicators of the multiple regression model used to examine the effect of selected financial indicators on tax debt across the 10 observed companies. The multiple correlation coefficient of 0.9569 indicates a very strong positive correlation, demonstrating the model's robustness in predicting tax debt. The R-squared value of 0.916 means that 91.6% of the variance in tax debt is explained by the included predictors, while the adjusted

R-squared of 0.874 confirms the model's high explanatory power, even after adjusting for the number of variables and sample size.

Table 5: ANOVA results – Model II

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	228.176.773,9651	76.058.924,6550	21,7250	0,0013
Residual	6	21.005.940,4998	3.500.990,0833		
Total	9	249.182.714,4649			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 80,0%*	*Upper 80,0%*
Intercept	4.075,9491	591,6916	6,8886	0,0005	2.628,1318	5.523,7664	3.224,0577	4.927,8406
koef. fin. stabilnosti	-6.636,3143	952,9834	-6,9637	0,0004	-8.968,1806	-4.304,4480	-8.008,3776	-5.264,2510
koef. obrta uk. imovine	2.072,2436	905,3939	2,2888	0,0620	-143,1755	4.287,6627	768,6975	3.375,7897
koef. zaduženosti	-2.341,6546	959,3842	-2,4408	0,0504	-4.689,1833	5,8740	-3.722,9336	-960,3757

Source: Authors' calculation

The ANOVA test ($F = 21.72$; $p = 0.0013$) confirms that the model as a whole is statistically significant, indicating that the combination of selected financial indicators meaningfully predicts tax debt. The financial stability ratio had a highly significant negative effect on tax debt ($\beta = -6{,}636.31$; $p < 0.001$), suggesting that firms with weaker long-term capital structures tend to accumulate more tax debt. This aligns with theoretical expectations, as lower financial stability often reflects insufficient solvency and a higher risk of delayed tax payments.

The asset turnover ratio showed a positive but marginally significant effect ($\beta = 2{,}072.24$; $p = 0.062$). Although it does not meet the strict 5% significance threshold, the result suggests a potential trend where firms with greater operational efficiency may still accumulate tax debt—possibly indicating that efficiency is not always linked to responsible tax behavior, or that other unaccounted factors may influence this relationship. The debt ratio also had a negative and borderline significant effect ($\beta = -2{,}341.65$; $p \approx 0.050$), implying that firms with higher levels of debt may face greater difficulty in fulfilling tax obligations. This may reflect financial pressure affecting tax discipline.

4.4. Discussion of the results

The results of this study reflect the essence of the KDD process, where structured financial data is transformed into insights with practical value (Fayyad et al., 1996). Using regression analysis, potential indicators of tax non-compliance were extracted from company-level financial indicators.

Model I confirms a significant negative relationship between EPS and tax debt, suggesting that more profitable companies are less likely to accumulate tax liabilities which is consistent with previous findings linking profitability to better tax compliance (Gupta & George, 2016).

Model II further shows that financial instability is a strong predictor of tax debt, which is consistent with theories that weaker capital structures reduce a firm's ability to meet its obligations (Škaro, 2024). The asset turnover ratio shows a marginally positive effect, suggesting that operational efficiency is not always correlated with fiscal discipline - possibly due to poor internal controls or liability management (Šarlija et al., 2009). The debt ratio also indicates a marginally negative effect, implying that higher indebtedness may limit a firm's flexibility in managing tax liabilities, reflecting broader findings on financial pressure and compliance behavior (Barić, 2022).

Overall, the findings support the idea that data mining within the KDD framework can be effectively used to identify tax risk patterns in financial data (Wahab & Bakar, 2021), providing a basis for more advanced predictive models in future research and potential early warning tools in public finance systems.

5. Conclusion

The challenges of managing large volumes of data are increasingly prominent in today's business and regulatory environment, making big data analytics and data science essential tools for informed decision-making. Data science, as a cornerstone of modern data-driven understanding, finds its technical and methodological expression in the Knowledge Discovery in Databases process. The KDD process represents a systematic, multi-phase methodological framework for transforming raw data into useful and applicable knowledge. It encompasses the collection, preprocessing, analysis, and practical application of data, while also facing numerous challenges such as data dynamism, heterogeneity, incompleteness, scalability issues, and the need for more efficient algorithms.

Through a comprehensive review of data mining techniques, including classification, clustering, regression, and anomaly detection, this study emphasizes their ability to extract meaningful and understandable knowledge from massive and diverse datasets. Moreover, the practical applicability of these techniques underscores their potential to uncover patterns and

predict behavior across various domains. The theoretical framework was further expanded by incorporating the concept of tax (non)compliance. Specifically, the use of regression models to predict tax debt based on financial indicators demonstrated the relevance of data mining techniques in identifying irregularities. This highlights that a systematic understanding and application of the KDD process requires not only technical expertise but also deep domain knowledge and an understanding of organizational dynamics. The analytical results of this study point to the significant role of factors such as earnings per share and financial stability in predicting tax liabilities, suggesting that financial statement data can serve as an early indicator for identifying high-risk taxpayers. Furthermore, the importance of integrating advanced analytics into decision-making processes is emphasized, along with the need for strategic development of analytical capabilities to enhance organizational efficiency and tax compliance. Finaly, the theoretical foundation presented in this research offers a strong basis for further investigation, development, and implementation of advanced analytical systems across all sectors of society – from public administration to the private sector.

The small number of observations represents a key methodological limitation. Due to the extremely limited sample of only ten companies, the results of this analysis should be considered exclusively indicative, since such a sample size seriously impairs the statistical power, stability of the model and the possibility of generalizing the findings. This paper presents a pilot study that demonstrates the possibilities of applying the KDD approach in detecting fiscal misalignment, with the findings being indicative and requiring additional empirical verification on larger samples. Hence, it is recommended that future research be conducted on larger datasets to validate these findings and incorporate additional variables that could help explain the dynamics of tax debt accumulation.

References

Agrawal, S., & Agrawal, J. (2015). Survey on anomaly detection using data mining techniques. Procedia Computer Science 60, pp. 708–713.

Agyei, S. K., & Yankey, B. (2019). Environmental reporting practices and performance of timber firms in Ghana: Perceptions of practitioners. Journal of Accounting in Emerging Economies 9(2), pp. 268–286.

Allingham, M. G., & Sandmo, A. (1972). Income tax evasion: A theoretical analysis. Taxation: Critical Perspectives on the World Economy 1(3-4), pp. 323–338.

Andreoni, J., Erard, B., & Feinstein, J. (1998). Tax compliance. Journal of Economic Literature, 36(2), pp. 818–860.

Arora, R., Pahwa, P., & Bansal, S. (2009). Alliance rules for data warehouse cleansing. In 2009 International Conference on Signal Processing Systems (pp. 743–747). IEEE.

Barić, B. (2022). Uloga financijske analize u planiranju poslovanja. (Završni rad, Sveučilište Josipa Jurja Strossmayera u Osijeku, Ekonomski fakultet)

Becker, M., & Buchkremer, R. (2019). A practical process mining approach for compliance management. Journal of Financial Regulation and Compliance 27(4), pp. 464–478.

Berkhin, P. (2006). A survey of clustering data mining techniques. In: Kogan, J., Nicholas, C., Teboulle, M. (Eds.), Grouping Multidimensional Data: Recent Advances in Clustering (pp. 25–71). Springer.

Boateng, K., Omane-Antwi, K. B., & Queku, Y. N. (2022). Tax risk assessment, financial constraints and tax compliance: A bibliometric analysis. Cogent Business & Management 9(1), 2150117.

Brady, H. E. (2019). The challenge of big data and data science. Annual Review of Political Science 22, pp. 297–323.

Cambridge University Press. (n.d.). Cambridge International Dictionary of English. Cambridge, NY: Cambridge University Press.

Cao, L. (2017). Data science: A comprehensive overview. ACM Computing Surveys (CSUR) 50(3), pp. 1–42.

Cao, L., & Zhang, C. (2007). The evolution of KDD: Towards domain-driven data mining. International Journal of Pattern Recognition and Artificial Intelligence 21(4), pp. 677–692.

Dedić, N., & Stanier, C. (2017). Towards differentiating business intelligence, big data, data analytics and knowledge discovery. In Piazolo, F., Geist, V., Brehm, L., Schmidt, R. (Eds.), Innovations in Enterprise Information Systems Management and Engineering. ERP Future 2016. Lecture Notes in Business Information Processing, (Vol. 285, pp. 114–122). Springer.

EMC Education Services. (2012). Data Science and Big Data Analytics. John Wiley & Sons.

Faizan, M., Zuhairi, M. F., Ismail, S., & Sultan, S. (2020). Applications of clustering techniques in data mining: A comparative study. International Journal of Advanced Computer Science and Applications 11(12), pp. 146-153.

Fan, C., Chen, M., Wang, X., Wang, J., & Huang, B. (2021). A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data. Frontiers in Energy Research 9, 652801.

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), pp. 27–34.

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). Introduction to KDD and Data science. In Learning from Imbalanced Data Sets (pp. 1–17). Springer.

Frawley, W. J., Piatetsky-Shapiro, G., & Matheus, C. J. (1992). Knowledge discovery in databases: An overview. AI Magazine 13(3), pp. 57–70.

Gama, J., Aguilar-Ruiz, J., & Klinkenberg, R. (2008). Knowledge discovery from data streams. Intelligent Data Analysis 12(3), pp. 251–252.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management 35(2), pp. 137–144.

Gorade, S. M., Deo, A., & Purohit, P. (2017). A study of some data mining classification techniques. International Research Journal of Engineering and Technology, 4(4), pp. 3112–3115.

Gullo, F. (2015). From patterns in data to knowledge discovery: What data mining can do. Physics Procedia 62, pp. 18–22.

Gupta, M., & George, J. F. (2016). Toward the development of a big data analytics capability. Information & Management 53(8), pp. 1049–1064.

Gupta, M., & Nagadevara, V. (2007). Audit selection strategy for improving tax compliance: Application of data mining techniques. In Proceedings of the Eleventh International Conference on e-Governance (pp. 28–30). Citeseer.

Han, J., Kamber, M., & Pei, J. (2012). Introduction. In Data Mining: Concepts and techniques, (3rd ed.). (pp. 1–38). Boston: Morgan Kaufmann.

Han, J., Pei, J., & Tong, H. (2022). Data mining: Concepts and techniques. Morgan Kaufmann.

Hernández-Blanco, A., Herrera-Flores, B., Tomás, D., & Navarro-Colorado, B. (2019). A systematic review of deep learning approaches to educational data mining. Complexity 2019, 1306039.

Hotho, A., Nürnberger, A., & Paaß, G. (2005). A brief survey of text mining. Journal for Language Technology and Computational Linguistics 20(1), pp. 19–62.

Houari, R., Bounceur, A., Kechadi, M.-T., Tari, A.-K., & Euler, R. (2016). Dimensionality reduction in data mining: A Copula approach. Expert Systems with Applications 64, pp. 247–260.

Igual, L., & Seguí, S. (2024). Introduction to Data Science: A Python Approach to Concepts, Techniques and Applications. Springer.

James, S., & Alley, C. (2002). Tax compliance, self-assessment and tax administration. MPRA paper No. 26906.

Jancsics, D., Espinosa, S., & Carlos, J. (2023). Organizational noncompliance: An interdisciplinary review of social and organizational factors. Management Review Quarterly 73, pp. 1273–1301.

Kamm, S., Jazdi, N., & Weyrich, M. (2021). Knowledge discovery in heterogeneous and unstructured data of Industry 4.0 systems: Challenges and approaches. Procedia CIRP 104, pp. 975–980.

Kirchler, E., & Wahl, I. (2010). Tax Compliance Inventory: TAX-I Voluntary tax compliance, enforced tax compliance, tax avoidance, and tax evasion. Journal of Economic Psychology 31(3), pp. 331–346.

Kubick, W. R. (2012). Big data, information and meaning. Applied Clinical Trials 21(2), pp. 26-28.

Lomotey, R. K., & Deters, R. (2014). Towards knowledge discovery in big data. In 2014 IEEE 8th International Symposium on Service Oriented System Engineering (pp. 181–191). IEEE.

Mikalef, P., Pappas, I. O., Krogstie, J., & Giannakos, M. (2018). Big data analytics capabilities: A systematic literature review and research agenda. Information Systems and e-Business Management 16, pp. 547–578.

Myles, G. D. (2000). Wasteful government, tax evasion, and the provision of public goods. European Journal of Political Economy 16(1), pp. 51–74.

OECD. (2018). Tax Challenges Arising from Digitalisation – Interim Report 2018: Inclusive Framework on BEPS, OECD/G20 Base Erosion and Profit Shifting Project. Paris: OECD Publishing. DOI: https://doi.org/10.1787/9789264293083-en.

Oza, N. C. (2008). Ensemble Data Mining Methods. In J. Wang (Ed.), Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications (pp. 356-363). IGI Global Scientific Publishing.

Pazzani, M. J. (2000). Knowledge discovery from data? IEEE Intelligent Systems and Their Applications 15(6), pp. 10–12.

Phalgune, A., Kissinger, C., Burnett, M., Cook, C., Beckwith, L., & Ruthruff, J. R. (2005). Garbage in, garbage out? An empirical look at oracle mistakes by end-user programmers. In IEEE Symposium on Visual Languages and Human-Centric Computing (pp. 45–52). IEEE.

Piatetsky-Shapiro, G. (1990). Knowledge discovery in real databases: A report on the IJCAI-89 workshop. AI Magazine 11(5), pp. 68–70.

Provost, F., & Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. Big Data 1(1), pp. 51–59.

Rahman, F. A., Desa, M. I., Wibowo, A., & Haris, N. A. (2014). Knowledge discovery database (KDD) – Data mining application in transportation. Proceeding of the Electrical Engineering Computer Science and Informatics 1, pp. 116–119.

Ristoski, P., & Paulheim, H. (2016). Semantic Web in data mining and knowledge discovery: A comprehensive survey. Journal of Web Semantics 36, pp. 1–22.

Roth, J. A., & Scholz, J. T. (1989). Taxpayer Compliance, Volume 1: An Agenda for Research. University of Pennsylvania Press.

Sagiroglu, S., & Sinanc, D. (2013). Big data: A review. In 2013 International Conference on Collaboration Technologies and Systems (CTS) (pp. 42–47). IEEE.

Santos, R. A. d., Guevara, A. J. d. H., Amorim, M. C. S., & Ferraz-Neto, B.-H. (2012). Compliance and leadership: The susceptibility of leaders to the risk of corruption in organizations. Einstein (São Paulo) 10(1), pp. 1–10.

Sebt, M. V., Sadati-Keneti, Y., Rahbari, M., Gholipour, Z., & Mehri, H. (2024). Regression method in data mining: A systematic literature review. Archives of Computational Methods in Engineering 31, pp. 3515–3534.

Seidu, B.A., Queku, Y.N., & Carsamer, E. (2023). Financial constraints and tax planning activity: Empirical evidence from Ghanaian banking sector. Journal of Economic and Administrative Sciences, 39(4), pp. 1063–1087.

Šarlija, N., Penavin, S., & Harc, M. (2009). Predviđanje nelikvidnosti poduzeća u Hrvatskoj. Zbornik Ekonomskog Fakulteta u Zagrebu 7(2), pp. 21–36.

Škaro, J. (2024). Likvidnost i financijska stabilnost Atlantic Grupe d.d. (Undergraduiate thesis, University of Rijeka, Faculty of Tourism and Hospitality Management).

Tang, J., & Karim, K. E. (2019). Financial fraud detection and big data analytics – Implications on auditors' use of fraud brainstorming session. Managerial Auditing Journal 34(3), pp. 324–337.

Usai, A., Pironti, M., Mital, M., & Aouina Mejri, C. (2018). Knowledge discovery out of text data: A systematic review via text mining. Journal of Knowledge Management 22(7), pp. 1471–1488.

Wahab, R., & Bakar, A. (2021). Digital economy tax compliance model in Malaysia using machine learning approach. Sains Malaysiana, 50(7), pp. 2059–2077.

Wamba, S. F., Gunasekaran, A., Akter, S., Ren, S. J.-F., Dubey, R., & Childe, S. J. (2017). Big data analytics and firm performance: Effects of dynamic capabilities. Journal of Business Research 70, pp. 356–365.

Wang, H., & Wang, S. (2008). A knowledge management approach to data mining process for business intelligence. Industrial Management & Data Systems 108(5), pp. 622–634.

Wu, R.-S., Ou, C.-S., Lin, H.-Y., Chang, S.-I., & Yen, D. C. (2012). Using data mining technique to enhance tax evasion detection performance. Expert Systems with Applications, 39(10), pp. 8769–8777.

Zandi, G., Obid, S., Hasan, R., & Ruhoma, A. (2019). Towards developing IT-based tax fraud detection models – The need for reform in the tax audit/investigation process. International Journal of Innovation, Creativity and Change 7(5), pp. 212–227.

Zemmouri, E. M., Behja, H., Marzak, A., & Trousse, B. (2012). Ontology-based knowledge model for multi-view KDD process. International Journal of Mobile Computing and Multimedia Communications (IJMCMC) 4(3), pp. 21–33.

Zhang, S., & Wu, X. (2011). Fundamentals of association rules in data mining and knowledge discovery. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1(2), pp. 97–116.

Zhou, Z.-H. (2014). *Ensemble Methods: Foundations and Algorithms*. New York: Chapman and Hall/CRC.

Sažetak

U savremenom poslovnom okruženju, analitika velikih podataka i tehnike rudarenja podataka sve se više prepoznaju kao alati za unapređenje fiskalne discipline i efikasnije upravljanje javnim prihodima. Ovaj rad istražuje mogućnost primjene procesa otkrivanja znanja iz baza podataka u otkrivanju obrazaca finansijskog ponašanja koji mogu ukazivati na poreznu neusklađenost. Korišten je kvantitativni pristup baziran na analizi sekundarnih podataka deset dioničkih društava iz Federacije BiH, za koja su dostupni finansijski izvještaji i podaci o poreznom dugu. Primjenom deskriptivne statistike i regresione analize ispitivana je povezanost između ključnih finansijskih pokazatelja (EPS, koeficijent finansijske stabilnosti, koeficijent obrta ukupne imovine i koeficijent zaduženosti) i iznosa poreznog duga. Rezultati pokazuju da niža profitabilnost i lošija finansijska stabilnost značajno koreliraju s višim iznosom poreznog duga, dok visoka operativna efikasnost i zaduženost imaju složeniji i statistički graničan uticaj. Nalazi potvrđuju mogućnost korištenja javno dostupnih finansijskih podataka za ranu identifikaciju rizičnih poreznih obveznika, što otvara prostor za dalji razvoj prediktivnih modela u domenu porezne analitike.

Ključne riječi: *porezna neusklađenost, porezni dug, rudarenje podataka, finansijska stabilnost, dobit po dionici, zaduženost, obrt imovine, regresiona analiza, KDD proces*