

нениям в общественных трансформациях, в переходах от одних способов производства к другим.

Изучение хозяйственных условий в органической связи интересами людей, что уже в исходной предпосылке предполагает исследовательскую линию, в которой постоянно присутствует взаимодействие всех сфер жизни человека, производственно-экономических, политических, социально-культурных процессов.

Экономика представляет собой науку об общих законах рационального человеческого поведения по сравнению с естественными науками, изучающими движение материи. Этот характер экономических законов Дж.С. Милль объяснял их сводимостью к «элементарным законам человеческого ума» [4, 293], иначе говоря, наличием некоторой общей природы человека, движимой в своём развитии интересами.

Принципиально значимым признаком противоречий в общественной жизни выступает столкновение интересов. Именно они находятся в основе переходов от одних способов производства к другим, приобретая новые формы выражения. Например, наиболее характерной чертой современности выступают противоречия между национальными и глобальными интересами, которые в тоже время не антагонистичны в той мере как классовые или политические.

Важнейшим моментом характеристики интересов как элементарных частиц в общественных процессах должна быть их сводимость, то есть обеспечение развития, ориентированного на реализацию совпадающих интересов. Совпадение интересов не только в индивидуальных, но и корпоративных, межстрановых масштабах создаёт предпосылки стабильного и гармоничного социально-экономического развития. Поэтому определение интересов людей как элементарных частиц в общественных процессах исходит из их возможностей понять "целое". В силу этого они в полной мере могут определяться как элементарные частицы не только в экономических, но и в общественных процессах.

Список литературы

1. Эйнштейн А. О методе теоретической физики. Собр. науч. трудов. Т.4 М.: Наука, 1965-1967.
2. Гэлбрейт Джон К. Экономические теории и цели общества. М.: Прогресс, 1976.
3. Гегель Г.В.Ф. Работы разных лет. т. 2. М.: Мысль, 1973.
4. Mill J.S. On the definition of political economy; and on the method of Investigation proper to it // Essays on Some Unsettled Questions of Political Economy. L.: 1844.

COMPARATIVE ANALYSIS OF OPEN-SOURCE DATA MINING TOOLS

Karalić I.,

Master engineer of organizational sciences

Pantelić O.

Associate Professor, Faculty of Organizational Sciences

Đukić M.

Teaching Associate, Faculty of Organizational Sciences

DOI: [10.5281/zenodo.7607428](https://doi.org/10.5281/zenodo.7607428)

Abstract

Companies strive to draw valuable conclusions from large amounts of data. These conclusions help improve decision-making, predict future trends, and gain a comparative advantage for that company. Mentioned processes can be provided using the data mining methods. Through data analysis, data mining indicates the connections between the data and provides valuable information. Numerous data mining tools are present on the market, some of them being free and others having to be purchased.

This paper analyzes three open-source data mining software products: Knime, RapidMiner, and H2O. It aims to contribute to a more straightforward decision-making process when choosing the data mining tool that meets the company's needs for quality analysis.

Keywords: comparative analysis, open-source software, data mining, Knime, RapidMiner, H2O.

Introduction

Responding to changes promptly is one of the critical actions that can provide a company with a competitive advantage in today's market, which is undergoing constant development and change. Having accurate and suitable data is as essential as the quality of the interpretation of the data. With changes in the market, data represents a significant resource for making quality decisions. The increasing amount of data collected by companies further aggravates the decision-making process. One of the challenges in the field of information technologies is reducing the time required for decision-making, that is, reducing the

uncertainty of achieving results. One of the techniques that arose in response to these challenges is data mining.

Data mining is a systematic and iterative data analysis process that enables better business decision-making. It can be defined as an automatic search for "hidden" information in databases [10]. "Hidden" refers to the connection of data. Based on those connections, a trend is observed, and a behavioral model is created.

According to [6], "data mining" is searching large amounts of data to discover patterns beyond simple analysis. It uses sophisticated mathematical algorithms

to segment the data to predict the probability of future events based on events that happened in the past.

Therefore, by applying this technique, behavioral matrices are created and used to predict future behavior, understand what is relevant for the company, and use it to achieve positive results. Data mining aims to transform large amounts of data into valuable information for companies.

The analogy with mining is quite apparent. In search of the precious ore hidden in the mountain, digging deep and throwing out large amounts of soil and stone is necessary. But once a vein is encountered, it must be followed along its entire length [3].

Data mining is based on three disciplines: statistics, machine learning, and artificial intelligence. These three disciplines define "data mining" as a process that studies and interprets data using algorithms that learn from the data and predict future behavior using software and statistical methods [11].

The data mining technique can be applied in various processes, such as production, banking, telecommunications, insurance, education, etc. Concrete examples of processes where data mining can help companies better understand data, per [2] and [9], are: acquisition of new and retention of existing customers; service improvement by analyzing data related to customer behavior regarding service, price, and distribution; risk management; improvement of relations with clients by predicting future behavior; examination of the best way of selling; and creation of customer profiles and segmentation.

This paper examines three open-source data mining tools that, according to Gartner, tend to become leaders in the field of data mining: Knime, RapidMiner, and H2O. These tools have been rated as visionaries for two years (2020 and 2021), indicating that they understand how the market is evolving and could be a good choice for leaders. It aims to study the characteristics of these tools to discover the most valuable tool for decision-making needs in different domains. In addition, the complexity of the learning-to-use process is also observed, as it can be significant factor in the choice. An analysis of the tools themselves and the application of these tools will contribute to understanding and identifying the best solution for a specific application domain.

Comparative analysis of Knime, RapidMiner and H2O tools

Considering the complex situation in the market, where it is increasingly challenging to respond to users' needs and stand out from the competition, extracting information from data is an important assignment. Information is a significant resource that a company

can use to maintain or improve its position. To provide a quality response to these challenges, the company must not only respond to the dynamic environment but also do so in the shortest time possible. Due to the large amounts of data that companies can collect, the need for quick and accurate solutions has outstripped the possibility of manual data analysis. That's why a company needs to find the right software that will enable it to quickly conclude on what decisions will be made. Aside from gathering the necessary data, the company may be required to test the software product and thus determine whether the observed product is the right solution.

When making decisions about which software product best suits the company's needs, the features and functionalities of the product must be examined in detail, as well as the installation possibilities, customizations, costs, and ease of use and learning. Also, it is necessary to determine whether there are certain use restrictions. Based on the information gathered and detailed analysis, the company can see the possibility of implementing the product and its usefulness in different aspects and thus make a timely and quality decision.

Knime, RapidMiner, and H2O are software products that many companies have already implemented, and Gartner rated them as possible leaders in data mining tools. The mentioned products are open-source tools allowing the developer community to use, upgrade, and share the software code freely. In this way, anyone who has programming knowledge and interest can contribute to building open-source software. However, although the code is publicly available, use and modification are protected by a software license that complies with the definition of open source given by the Open Source Initiative [12]. Open source products have numerous advantages, some of which are [14]: the code is free and available for connection and integration with other products; the possibility of adapting the product to its environment; no hidden costs; by adopting an open source product, the company does not depend directly on the manufacturer; and the possibility to upgrade the product at any time.

Technology and operating system

Figure 1 shows the essential technical characteristics of observed open-source software products for data mining (Knime, RapidMiner, and H2O). If some product does not provide versions for specific operating system that the company uses, it is clear that this product cannot be a suitable solution for the company.

Figure 1

Versions of observed products for different operating systems

Product / Operating system	Knime	RapidMiner	H2O
Windows	yes	yes	yes
Mac OS	yes	yes	yes
Linux	yes	yes	yes
Technology	Java	Java	Java

It can be noted that for each product, there is a version corresponding to Windows, MAC, and Linux operating systems. Also, all three products are based on Java technology. It implies that companies that opt for these products should have employees who possess knowledge in the domain of Java.

Software installation

As for the installation of these products, the process itself is pretty simple. The prerequisite for Knime and RapidMiner installation is registration. Furthermore, during product startup, RapidMiner requires login to the created account. Installation of the H2O product does not require any registration or login.

Implementation and maintenance costs

When choosing the right product, the cost of using it is an important factor. Low prices characterize open-source software products, and there are opportunities to use the software for free. When it comes to costs, they

can include the cost of acquisition, implementation, maintenance, and product customization.

When it comes to open-source Knime, RapidMiner, and H2O products, all three offer the possibility of free use. Knime and RapidMiner also offer solutions with more functionalities intended mostly for larger companies. These solutions require payment.

Knime offers one version of the product, the Knime Analytics Platform, which is free for usage. Also, Knime offers an additional product, Knime Server, which is a supplement to the free version of the product for companies that need to share information between teams, better control of adherence to the data protection policy, automatic flow planning, etc. [4]. Knime offers three versions of the Knime Server platform: Knime Server Small, Knime Server Medium, and Knime Server Large.

Figure 2

Characteristics of Knime Server solutions [4]

Functionalities	Knime Small	Knime Medium	Knime Large
Purpose	For small teams	For middle size teams	For large teams
Collaboration			
Share workflows and control access rights	yes	yes	yes
Upload and share components to enable users to reuse most common functionalities	yes	yes	yes
Customize the node repository to ease use and ensure compliance			yes
Automation			
Schedule a workflow or report to be executed at a certain time, or periodically	yes	yes	yes
Use Workflow Pinning for automated routing of workflows			yes
Deployment			
Create and deploy Guided Analytics	yes	yes	yes
Deploy workflows via the REST API to allow access from other applications		yes	yes
Number of consumers with access to analytical applications	No free customers	Limited customers	Unlimited customers
Management			
Create workflow snapshots and compare to previous versions	yes	yes	yes
Monitor server activity (running and scheduled jobs), adjust permissions, manage ongoing services	yes	yes	yes
Access detailed summaries of workflows for data lineage		yes	yes
Integrate authentication with corporate Active Directory setups, and Single Sign-On			yes
Pricing			
Yearly	14.500 euros	25.000 euros	45.500 euros

RapidMiner offers three versions of the product: RapidMiner Studio Free, RapidMiner Studio Educational, and RapidMiner Studio Enterprise. RapidMiner Studio Free is free but has a limit on the amount of data it can analyze. Namely, this version of the software supports a maximum of 10,000 rows of data. So, if the company needs to analyze more than 10,000 rows of data, then this version of the software is not a good choice. Companies can also use the Educational version, which is also free. This version is limited in time, it is free for the first year. It offers every

functionality, so the company can opt for this version for the first year. For companies whose databases exceed 10,000 rows, RapidMiner offers an Enterprise version with free product usage during the first month. The cost of using this product is not publicly available; it is calculated based on the user's needs. The user can send an inquiry through the RapidMiner website.

As for the H2O product, this company does not highlight the cost of using the product. Online, it can be found that this expense can cost the company up to

\$300,000 for a three-year subscription or \$850,000 for a five-year subscription.

Customer support and ease of use

When choosing software, documentation, the possibility to contact vendors, blogs, and other sites where product users can find support are significant factors. In the first place, customer support allows users to understand the product faster and better via tutorials explaining the purpose and functionality. Also, various types of customer support can help users learn more quickly how to use the product and thus affect the simplicity of their use.

Knime offers a large selection of support documents. There is a separate section on the site intended for users who are familiar with the product. This section contains instructions regarding user education, explains the purpose of the Knime software, and gives an example of how to use the product. There is also documentation that describes the Knime product's interface and functionalities, as well as workflow components and their functionality. A blog containing various materials on the application of the Knime product, the certification program, and other data mining or analytics topics is also available to users. Knime offers installation support; documentation on the site covers the complete installation process. In addition to installation, guides can be found that explain best practices in usage, component, file management, integrations, etc. In addition to detailed documentation, Knime also has a forum that allows users to post specific questions and receive answers related to the process of using the software and any eventual difficulty they encounter.

RapidMiner, like Knime, provides user support in the form of documentation. Instructions regarding

installation can be found on the official site, as well as initial steps, connection of RapidMiner Studio with other applications, etc. It also provides guidance on interfaces, nodes, components, building flows, using functions, and visualization of results. Another form of support for users of the RapidMiner product are tutorials available during the first product startup. Through the tutorials, users can go through various cases of model-making, from simpler to more complex, and become acquainted with the features of RapidMiner. RapidMiner also has its community which provides support to RapidMiner users.

H2O provides customer support through documentation available on the official website for installation and basic use. In addition to the documentation, it also contains an option in the framework platform for help; that is, it offers more detailed explanations of functionality.

In terms of ease of use of the product, Knime and RapidMiner provide different forms of support that have an impact on the easier mastering and learning how to use the products. Apart from that, the user interface affects the ease of use. With Knime and RapidMiner, the user interface is entirely graphical, which allows users to use the product without programming knowledge. With the H2O product, the interface also requires text input for commands. Therefore, if the company wants to opt for H2O, it should keep in mind that this product's users must have basic programming knowledge.

Functionalities

Figure 3 shows the products described in this work and available data mining methods. It can be noted that H2O does not support the application of decision tree method or the application of association rules.

Figure 3

Data mining methods available across observed products

Data mining method	Type	Knime	RapidMiner	H2O
Neural nets	predictive	yes	yes	yes
Naïve Bayes	predictive	yes	yes	yes
Decision tree	predictive	yes	yes	no
Support vector machine	predictive	yes	yes	yes
Association rules	descriptive	yes	yes	no
Clustering	descriptive	yes	yes	yes

Churn prediction in a telecommunications company using RapidMiner

In telecommunications, data mining can help a company with several aspects [2]: acquisition of new users, during development (i.e., increasing the value for the user), and retention of existing users.

Various methodologies can be used to define the goal of the analysis, the phases in the process, data, tools, and expected results. In practice, methodologies usually contain four to eleven stages [1]. Regardless of the number of stages, the steps in these methodologies tend to be the same. Each methodology assumes that it is crucial to understand the problem, prepare the data, and then start building the model. Following construction, the model should be checked and validated, and any deviations should be corrected. After the model is checked, the same is applied to the data to obtain valuable information for decision-making.

Finally, it is necessary to interpret, present, and monitor the results.

In this paper, the analysis will be done according to the data mining methodology proposed by Zorica Bogdanović [7]. An open-source product, RapidMiner, will be used for this analysis because of its functionalities. The data used in the analysis is real data collected by a telecommunications company in Serbia.

Problem definition

As Peter Drucker said [16]: "The business of all businesses is to win and keep customers." It means that companies have a task to acquire and retain users, despite the increasingly complex market situation and demanding users. A satisfied user is a concept that gets more attention in theory and practice because every unsatisfied user will look for another solution.

The analysis should examine how specific parameters influenced users' decisions to terminate the

contract and predict the decisions of users who are still part of the company. Target variable is churn indicator that shows whether the user remains part of the company or has switched to the competition.

Research and data preparation

The data is recorded in an SAP product by the telecommunications company whose data is being analyzed. In this company, every contact with the user (interaction) is recorded, whether the user calls the contact center or visits the sales location, as well as interactions when the company contacts the customer. Based on the interactions, data is collected about the nature of the contact (e.g., information, complaint, technical support), the frequency of communication (e.g., whether the user calls every day or once per month), the length of the conversation, the time of establishing contact (e.g., the user mostly calls in the morning), the method of establishing a connection (e.g., the user mainly establishes contact in person or upon arrival at the sales location), etc.

One of the parameters that can be of great importance for this analysis is the net promoter score (NPS). NPS is a metric the observed company uses to assess to what extent its users are satisfied with the company. NPS gives users the ability to rate a company from 0 to 10, answering the following question: "On a scale of 0 to 10, how likely are you to recommend the company to your friends and family?" It is an important indicator because only highly satisfied customers confident in the company will recommend the company to their friends.

The following parameters will be taken into account: a place of service used; the number of interactions categorized as complaints; NPS indicator as an indicator of satisfaction; regularity of bill payment; a package of services used by the user; the total number of interactions (contacts with the company). Users from Belgrade, Novi Sad,

Kragujevac, Kruševac, Niš, Valjevo, Čačak, and Užice who became users in the second quarter of 2022 are included in the data.

Model building

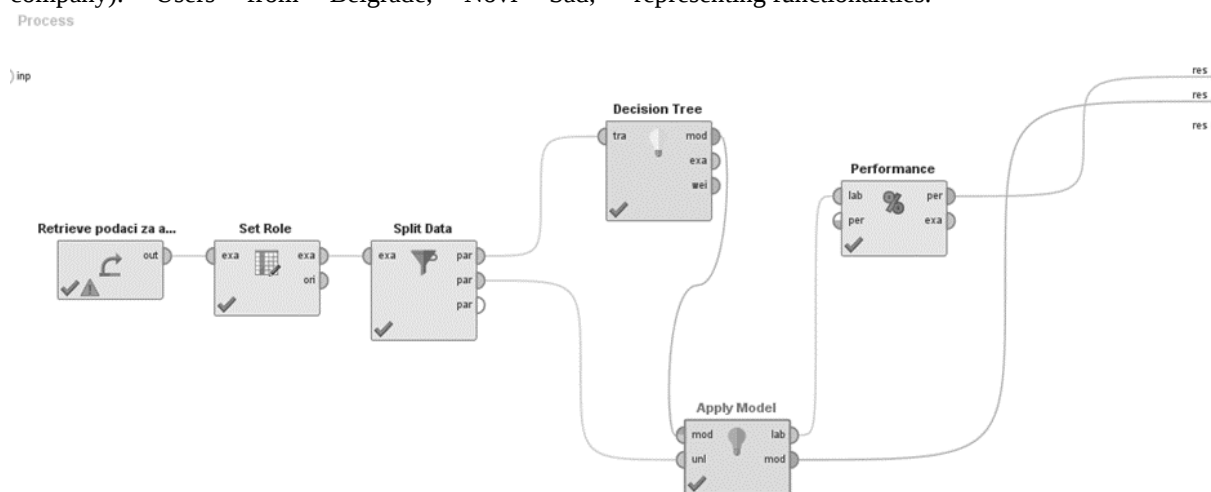
The decision tree method was used in this example to predict the results. The choice of method depends on the nature of the problem, so it cannot be determined in advance which method is the best choice for the problem. This decision is usually made by the person who is familiar with advantages and disadvantages of the method, e.g., a data mining expert.

Data mining methods can be divided into two groups [5]: predictive, which aims to confirm the hypothesis, and descriptive, which is oriented towards discovering patterns and looking for trends, but without prior knowledge of the target variable.

Descriptive methods include clustering and association rules. In clustering, variables are placed into categories based on their attributes' similarity, making this technique reminiscent of classification. The difference is that clusters are not known in advance. Association rules involve discovering trends and connections between two or more variables. The association rules method is often used to identify crucial parameters in the observed process and their mutual relations [13].

Decision tree is a predictive method, along Naïve Bayes, neural nets and support vector machine (SVM). These methods include training and test data. Based on training data, rules are observed. Those rules are further used on data whose outcomes are unknown (test data) to draw conclusions [8]. Another predictive technique that aims to analyze functional or stochastic interdependence between parameters is regression [15].

Picture 1 shows a built-in model for predicting user departure in the RapidMiner product. As this paper explains, the model consists of components representing functionalities.

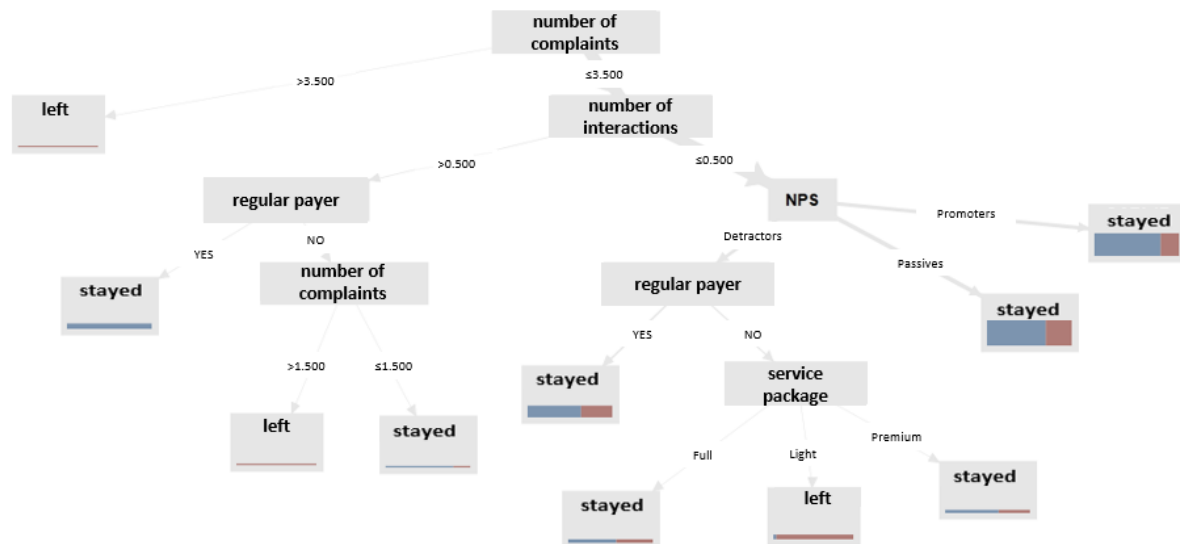


Picture 1 Churn prediction model

Since churn prediction is a classification problem, the task of the model is to predict, in relation to the previously explained parameters, whether the user is a member of the class that "remains" a part of the company or "goes" to the competition.

Analysis results

The result of the analysis is presented as a tree [Picture 2], since the decision tree technique was used.



Picture 2 Analysis results

The results show that in the observed sample, the number of reported complaints is the parameter that significantly influences the user's decision to remain a user. Also, the results show that users with more than four complaints, regardless of other parameters, decided to switch to the competition. For the company, it may mean that the focus in the process of retaining users should be precisely the quality of the infrastructure and the quality of the signal, to reduce the number of complaints.

For users with fewer than four complaints, it can be noted that the decision depends on the total number of interactions. For users who have at least one interaction recorded and are also paying their bills regularly, it can be concluded that they will decide to stay. However, if they are not paying regularly and have more than one reported complaint, they will choose to terminate the contract. It also indicates that the number of complaints is decisive in the user's decision. If there is no interaction from the user, the NPS parameter may be necessary. According to the findings, detractors who use the Light package of services and do not pay their bills regularly are more likely to switch to the competition.

Therefore, based on the analysis results, it can be concluded that the company needs to devote attention to users who have registered complaints, as well as to users who do not pay bills regularly. According to this analysis, the company can define different activities (within the company and towards the users) that will influence the users' decision to stay a competition.

Conclusion

Constant changes in the environment require continuous development and further complicate the decision-making process for companies. Companies must use all their resources to maintain and improve their position in the market. One of the most important resources today is information. It is of great importance that the information based on which decisions are made be accurate and timely. Data mining is a tool that allows businesses to make quality decisions based on the data they collect.

Since data mining methods are widely used in various industries (banking, finance, marketing, sales, healthcare, production, education, telecommunications, etc.), numerous companies have tools that help them understand the environment.

Knime, RapidMiner, and H2O represent products that, according to Gartner, tend to become leaders in the field of data mining. Through the comparative analysis of these products, it was observed that all three are available on Windows, Linux, and MAC operating systems and are based on Java technology.

The difference is observed in the functionalities. Knime and RapidMiner have an advantage since they support a more significant number of methods that can be used in the analysis. From a customer support point of view, these two products provide numerous documents publicly available to users through the website, blog, and forum. In addition to documentation, they provide tutorials for beginners. Based on this, it can be concluded that Knime and RapidMiner are also suitable for users who do not have a lot of knowledge about data mining and methods, but also for users dealing with more complex problems due to their functionalities.

H2O is a suitable solution for users dealing with less complex or specific problems who prefer coding and have appropriate knowledge.

All three products offer free versions, with the free version of RapidMiner being limited either by time or amount of data; therefore, for users who need to analyze more than 10,000 rows of data, RapidMiner Studio is not the best solution. These users can opt for Knime.

In further work, it is possible to examine in more detail additional products offered by these vendors, for example, Knime Server and RapidMiner Enterprise, and include them in the analysis. Also, the focus of this paper was an analysis of open-source data mining products; however, there are also proprietary products with numerous advantages; therefore, the work could be extended by analyzing proprietary products and by comparing proprietary and open-source products.

References

1. D. M. M. Ali, Role of Data Mining in education sector, *International Journal of Computer Sciences and Mobile Computing*, p. 374-383, 2013.
2. D. Q. B. A. Kazi Imran Moin, Use of Data Mining in Banking, *International Journal of Engineering Research and Applications (IJERA)*, pp. 738-742, 2012.
3. I. Čabrilo, Iskopavanje podataka, 01.05.2005. Available at: www.sk.rs/2005/05/skpr01.html. Accessed: 10.07.2022.
4. Knime, KNIME Server Pricing, 12.07.2018. Available at: www.knime.com/knime-software/knime-server-pricing. Accessed: 13.08.2022.
5. L. B. L. G. G. L. a. J. d. O. Luis Martín, Using data mining techniques to road safety improvement in Spanish roads, *Procedia - Social and Behavioral Sciences*, p. 607 – 614, 2014.
6. M. B. ENEROTH, An analysis of customer retention using data mining, *KTH ROYAL INSTITUTE OF TECHNOLOGY*, t. 33, p. 7, 2018.
7. M. D. B. R. Zorica Bogdanović, DATA MINING U SISTEMU ELEKTRONSKOG OBRAZOVANJA, *INFO M*, pp. 26-34, 2006.
8. V. G. Fougatsaro , A Study of Open Source ERP Products, *SCHOOL OF MANAGEMENT BLEKINGE INSTITUTE OF TECHNOLOGY*, Karlskrona, 2009.
9. N. J. Prof. Dr Milan Milosavljević, *Alati za Data Mining*, Beograd, 2011.
10. P. d. K. S. Prof. dr Margarita Janeska, *Data mining - Put ka konkurentnosti*, Ekonomski fakultet - Prilep, Prilep, 2005.
11. SAS, Data Mining: What it is & why it matters, 20 05 2015. Available at: www.sas.com/en_us/insights/analytics/data-mining.html. Accessed: 10.07.2022.
12. Bjeladinović, S. (2018). Materijali sa predavanja iz Integrisanih softverskih rešenja. Univerzitet u Beogradu, Fakultet organizacionih nauka, Beograd, Srbija.
13. S. Milinković, *Korišćenje asocijativnih pravila za istraživanje edukacionih podataka*, Sarajevo, 2015.
14. V. G. Fougatsaro , A Study of Open Source ERP Products, *SCHOOL OF MANAGEMENT BLEKINGE INSTITUTE OF TECHNOLOGY*, Karlskrona, 2009.
15. V. Petrovic, *Teorijske osnove za izradu master rada*, Univerzitet u Beogradu, Tehnički fakultet u Boru, Bor, 2016.
16. Z. Erdeljan, *ZADOVOLJSTVO KORISNIKA*, Portal kvalitet, 2017