International Journal of Engineering Inventions
e-ISSN: 2278-7461, p-ISSN: 2319-6491
Volume 13, Issue 4 [April. 2024] PP: 217-222

Data Mining - The Path to Advanced Higher Education

Srđan Mitrović[1], Muktar Mohamed Emhemdi Albueshi[1], Dejan Rančić[2], Filip Marković[3], Mladen Veinović[1]

[1] *Singidunum University Belgrade, REPUBLIC OF SERBIA*
[2] *Faculty of Electronics Niš, University of Niš, REPUBLIC OF SERBIA*
[3] *Faculty of Technical Sciences Kosovska Mitrovica, University of Priština in Kosovska Mitrovica, REPUBLIC OF SERBIA*
Corresponding Author: Filip Marković
e-mail: filip.markovic@pr.ac.rs

ABSTRACT: *The paper explores the wide application of Data Mining techniques in the context of higher education, with a special focus on the analysis of student data. Using methods such as classification, clustering, assessment and visualization, we explore how Data Mining can contribute to the improvement of the educational process. Through examples such as recommendations for choosing faculties and courses for future students, and predicting student success based on their academic performance, we illustrate the practical application of Data Mining in higher education. In this way, these techniques can help personalize the educational experience, providing students with tailored recommendations and support. Finally, we highlight the critical importance of Data Mining in the context of continuous improvement of the educational experience, providing insights that help students, teachers and institutions make informed decisions to achieve better learning outcomes.*

Keywords: *Data Mining, academic institutions, classification, clustering*

Date of Submission: 15-04-2024 Date of acceptance: 28-04-2024

I. INTRODUCTION

Data Mining can be defined in several ways that differ mainly in their focus. A very early definition of Data Mining was "the non-trivial extraction of implicit, previously unknown and potentially useful information from data" [1]. A later definition of Data Mining expanded this definition a bit, referring to the application of various algorithms to find patterns or relationships in a data set [2]. Therefore, knowledge discovery referred to the additional process of data access, data processing, data post-processing and result interpretation. This is a very useful approach for expressing all the steps needed to find and exploit relationships in data.

The high need for educated staff requires continuous improvement of the teaching process at higher education institutions. A quality educational process begins with an analysis of the target group of students and the future requirements of the industry they are pursuing, which becomes the basis for the accreditation of study programs. Monitoring generations of students during the course provides insight into their behavior, results and suggestions, which contributes to the improvement of the course itself.

The interest of higher education institutions is also directed towards predicting the trajectory of studies and the success of students. This includes assessing which courses students will choose, as well as predicting their success. Another important topic is the failure of students to complete their studies, which has sparked numerous debates.

Researchers Batool et al. [3] tried to classify students according to their future achievements, risks of failure in exams and assessment of their level of knowledge in order to detect deficiencies in the learning process. Such analyzes help to identify factors that can influence student failure and enable timely taking of appropriate measures to improve educational processes.

In today's digital age, education is increasingly relying on data analytics to enhance the learning experience. The data that students collect about institutions or courses becomes a key factor in the decision-making process about the future of education. Information coming from previous generations of students can provide valuable insight into the characteristics of courses and institutions, helping future students make informed decisions. Through proper processing of this data using Data Mining techniques, key patterns and trends can be identified, allowing universities to tailor their programs and resources to better suit student needs.

Data Mining techniques are now increasingly used in education to optimize the learning process and support student development. Through data analysis, academic institutions can predict future trends in student behavior, identify at-risk groups and adjust their strategies to improve learning outcomes. These tools also allow

institutions to personalize the learning experience and support students on an individual level.

The Data Mining process relies on four basic methods: classification, clustering, evaluation and visualization [4]. Through these methods, patterns in data are more deeply understood and used to make informed decisions in education. For example, data classification can help identify factors that influence student success, while clustering can group students according to their needs and characteristics.

II. DATA MINING IN EDUCATION

Originally, the definition of Data Mining was limited only to the process of building models. But as the practice matured, Data Mining toolkits included other necessary tools to facilitate data preparation and model evaluation and display.

Data Mining uses two basic models: descriptive models, which use unsupervised learning to identify patterns in data without target benchmarks, and predictive models, which typically use supervised learning to explain the values of a target variable relative to other variables. Data Mining tools use sophisticated analytical capabilities, knowledge bases, and problem domain understanding to uncover hidden trends and patterns among data, enabling analysts to draw new conclusions. This technology helps to discover valuable information and knowledge in large data sets, using supervised or unsupervised learning algorithms to analyze unstructured data [5].

As already mentioned, Data Mining is based on four key methods, each of which is used in different situations and contains its own sub-methods and algorithms.

Classification in its most basic form involves grouping data into certain classes or groups based on their similarity. To determine similarity, distance measures are commonly used. Two basic requirements for a successful classification are that the classes are exhaustive (all data can be assigned to one class) and mutually exclusive (each data can belong to only one class). Although there may be several classes to which no data is assigned, there must be no data that cannot be assigned to any class. When working with data in a database, classification is often used to establish a function that allows data to be sorted into one of several defined classes. Similarity is usually determined using distance measures. Classification uses predefined classes [6].

Clustering, as another large class of techniques in data analysis, aims to divide a given data set into homogeneous subsets. This technique is often called unsupervised because it does not require predefined classes or training examples [7].

In the context of Data Mining, data clustering is the process of determining groups of data that are similar to each other, but at the same time different from other data. When clustering, key variables that allow for the best grouping are often identified. The key assumption of this process is the clustering hypothesis, which claims that all documents in the same cluster share similar properties, on the basis of which they are assigned to that cluster, but also that they have the same relevance of the data found in the documents [8].

Visualization is a powerful tool for showing mathematical rules or research results. It can use interactive graphs to display data in an intuitive way. It was originally used to visualize three-dimensional geographic locations based on mathematical coordinates.

On the other hand, estimation involves the analysis of prediction functions or the probability of a particular event. These functions apply to variables that change continuously. Based on these assessments, key business decisions can be made that will further guide the institution's activities.

Data Mining is a key tool in modern business, enabling a deeper understanding of data and uncovering hidden patterns and information. This multidisciplinary field uses various techniques and algorithms to analyze large data sets and extract useful insights for informed decision making. One of the most commonly used frameworks in Data Mining is the CRISP-DM model, which defines the hierarchical structure of the process and enables a systematic approach to data analysis [9].

The CRISP-DM model, developed by leading companies such as NCR (National Cash Register Corporation), SPSS (Statistical Package for the Social Sciences) and Daimler-Benz, consists of several key stages: from defining business goals and understanding the context of the problem, through data collection and preparation, to modeling, evaluation of results and implementation in practice. These stages form the foundation of the Data Mining process, enabling organizations to strategically use their data to improve business processes and achieve competitive advantage [10].

Each stage of the CRISP-DM process is shown in Fig. 1. CRISP-DM represents a hierarchical model consisting of six basic phases. Starting with the definition of business goals, data collection, preparation and cleaning follows as the second phase. Then the process continues with modeling, evaluation of results and finally distribution of project results. Each of these phases is broken down into different scenarios, while at the third level tasks are specialized for specific situations. For example, one of the generic tasks might be cleaning data, while a specialized task might be about cleaning numeric or categorical values. At the fourth level, there is an instance of the process, which includes concrete actions, decisions and results of project implementation [9].

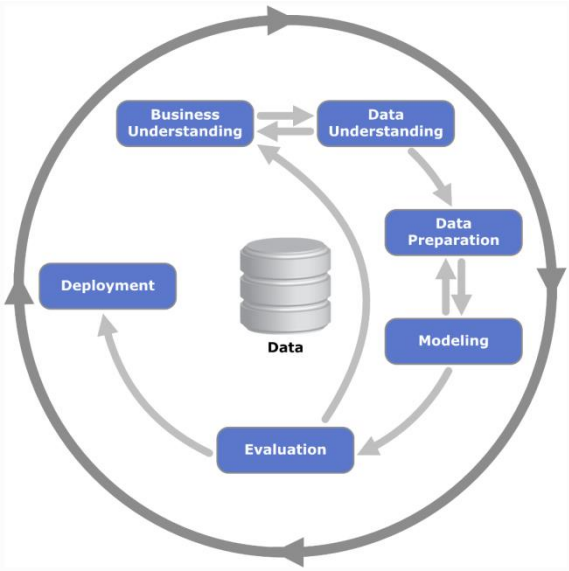

Fig. 1 Processesin Data Mining [11]

Modeling in the context of Data Mining not only considers the relationships between different tasks, but also provides an idealized sequence of actions over the life of the project. This structured methodology enables the organization and effective management of the process, providing a clear map of activities.

In the field of education, the application of the basic processes of Data Mining brings the possibility of using various techniques for data analysis. Among these techniques, decision trees, k-nearest, neighbor neural networks, and many others, offer deeper insights and understanding of hidden patterns in educational contexts. By using these techniques, it is possible to discover different types of knowledge that can be invaluable for improving the educational process. The applications are diverse: the organization of the curriculum, the prediction of the number of students who will enroll in a certain study program or choose a certain course, the identification of shortcomings in the traditional teaching method, the detection of irregularities in electronic testing, as well as the detection of anomalies and illogical results in student papers [12].

These techniques not only enable a more efficient organization of the educational process, but also provide valuable insights that can be essential for the continuous improvement of the quality of education for both students and teachers.

III. DATA MINING TECHNIQUES IN IMPROVING HIGHER EDUCATION: ANALYSIS, PREDICTION AND OPTIMIZATION

Data Mining in higher education is becoming an increasingly important research area, as it provides significant opportunities for the improvement of academic institutions [13].

In the continuation of the paper, examples of the application of various Data Mining techniques in the process of improvement, evaluation and development of various activities in the educational process are presented. These techniques can be applied from the beginning of student enrollment and the selection of the appropriate faculty, through monitoring the study process and providing support during studies, up to the assessment of students' success in graduation and the formulation of enrollment policies, both at the level of the faculty, as well as at the level of the university and society in general.

3.1 Application of Data Mining Techniques in the Faculty and Course Selection Process: Helping Decision Making in an Academic Career

The correct choice of faculty and study program has become a key step for the formation of a future career. The decision that any individual makes at the beginning of their academic career can have long-term consequences for their future. This selection process is complex and often requires good thought and research.

One of the key factors that affects the future occupation of each individual is the right choice of faculty. The choice of faculty should be aligned with the abilities, interests and goals of the future student. Each faculty has its own specifics and approach to education, therefore it is important for the student to find an institution that matches his needs and preferences. Another important step is choosing the appropriate study program within the chosen faculty. Different study programs provide different perspectives and opportunities for developing specific skills and knowledge. The student should carefully consider the content of the program, its alignment with his interests and future career goals. College curricula are often flexible, allowing students to shape their education according to their interests and goals. Choosing the right courses can help a student gain

in-depth knowledge in areas of particular interest and relevance to a future career.

In essence, the final decision often depends on experience and available information. Developing a system based on Data Mining techniques that could recommend appropriate courses to students during their studies, using data on previous generations of students who attended the same courses, could be extremely beneficial. The data that would be used would include demographic information, course selections, grades achieved, number of courses in the same field, average grades during the course of study, as well as average grades in courses in similar scientific fields. After collecting and processing data using the developed system, students would be offered recommendations on whether to choose a particular course and how likely they are to succeed in that course. In this way, students would have a useful tool to help them make an informed decision about whether or not to enroll in a particular course. In the process of developing such a system, the authors would first perform a data cleaning procedure in order to eliminate all data that are not suitable for further analysis. Then the classification and identification of patterns in the data of students who showed interest in the same course would be done. With the application of the C4.5 algorithm, the system was able to suggest to students which course they should choose with 80% accuracy [14].

In the college selection process, informed decision-making is essential for each student's future. The implementation of a system based on Data Mining provides effective support to this process. By analyzing various relevant data, this system enables the assessment of compatibility between a future student and a potential faculty. Key aspects taken into account include high school grades, vocational guidance, extracurricular activities and experiences of previous students with a similar profile. The goal of the system is to facilitate decision-making for high school graduates by providing objective information on the basis of which they can evaluate their options and choose a faculty that matches their interests and goals. Also an aspect that cannot be ignored are the experiences of people who come from the same school with the same or similar average grades and interests. In this way, it would be easier for high school graduates to decide which faculty to enroll in, and on the other hand, they would be presented with objective data on the basis of which the assessment was made. The measure of success of the system would be the number of students who would successfully enroll and graduate from the selected faculty.

3.2 Predicting Student Success: Application of Data Mining Techniques in Academic Performance Analysis

During the study process, different generations of students achieve different results. While some students are more successful in mastering the course, others need more attention to achieve the same level of success. This variability in results makes it difficult to accurately predict the success of completing a course or an entire degree program. In these situations, Data Mining techniques can be useful. By analyzing the data on the current group of students in a certain course, it is possible to estimate their probability of success in the final exam. This data is especially useful if there is a dependency between success in this course and enrollment in another course or program.

For this purpose, the k-means algorithm is used. The K-means algorithm is a method of clustering data on student activities. The K-means clustering algorithm is the most popular and well-known clustering technique. Among various clustering techniques, the k-means algorithm is widely used due to its simplicity, ease of use, and high performance. It can be concluded that the k-means method is highlighted as one of the most important and significant clustering approaches, especially in the context of predicting the academic performance of students [15].

The operation of the clustering algorithm consists of two steps. In the first step, the algorithm selects initial cluster centers. Then, in a second step, the algorithm moves those cluster centers to minimize the sum of the squared distances of all points from the centroid. This sum of squared distances is known as RSS (Residual Sum of Squares). After that, these steps are iteratively repeated until the condition for terminating the iterations is met, which includes reassigning the objects to clusters based on the nearest centroids. Before clustering is performed, the data to which this process is applied should be processed and placed into initial clusters [16]. This includes data on previous academic results, practical exams and course tests. Students are classified into three classes based on their results, and these three classes are formed based on the number of points earned during the course.

The first class represents a high number of points (up to 70 points), the second class is a medium number of points (from 50 to 60) and a low number of points (below 50). Based on this, students can be divided into two classes, namely the class of students who will pass the exam and the class of students who will not pass the exam. The success of completing a certain course, as well as the activities carried out during the course, depend on many factors and can be viewed from different aspects. The most important factors that separate successful from unsuccessful students can be divided into socio-demographic and learning environment factors [15]. Sociodemographic factors include: age, gender, ethnicity, education, work status, and disability, while learning environment factors include: course program and semester in which the course takes place. Regarding

the age of students, students can be people of different ages depending on the level of study. In general, they could be divided into students under thirty, between thirty and forty, and over forty. Ethnic diversity in academic institutions varies by location. Some universities may have more students from different ethnic groups, while others may have less. In addition, the number of students in each ethnic or national group has a significant impact on performance evaluation. For example, if the number of members of a certain ethnic group is small, that data can be ignored or grouped with similar groups for more accurate results. Also, it is important to emphasize the presence of students with disabilities, because this can affect their success. Previous education also plays a role, so it is useful to group students according to the level of education they have previously received. Also, the work status of students can be employed or unemployed, which can affect their ability to attend classes and fulfill their obligations. Classification trees can be applied to such data to identify factors that influence students' success in completing a course. Research has shown that classification trees, especially CART (Classification and Regression Tree), are useful, achieving a classification accuracy of 60.5%. However, although this classification may indicate a risk of failure, it is possible to take additional steps to support students who face challenges in completing the course, such as counselling, mentoring or setting specific goals.

3.3 Distance Learning: Application of Data Mining Techniques in Detecting Cheating in E-Learning

The development and application of new technologies in learning have led to the possibility of distance learning, which has become especially significant in the context of the COVID-19 pandemic. In such a system, students are not present in the premises of the faculty, but can have access to the content of the course and follow the lessons from anywhere. E-learning has grown significantly every day over the last decade with the growth of the Internet and technology. Therefore, an online exam can be useful for people who want to take the exam, but cheating on tests is a common phenomenon all over the world. As a consequence, cheating prevention can no longer be fully effective. The electronic learning system enables the use of Data Mining techniques both in the process of organizing classes and gathering information to improve the educational process, as well as in the field of testing and final exams. This technology enables the analysis of large amounts of data to identify learning patterns, personalize instruction according to students' needs, and improve the effectiveness of evaluation and assessment of their success. In a situation where physical presence at the university is limited, distance learning becomes a key resource for maintaining the continuity of education and supporting students in achieving their academic goals [17].

The organization of taking exams via electronic portals requires adequate measures to prevent abuses, identify cheating and correctly assess responsibility for the questions asked. When it comes to cheating on such portals, the responsibility lies not only with the students, but also with the exam organizers. Data Mining techniques can help detect and prevent cheating in online systems, using data on known cheating methods to achieve more effective results. Another way to detect irregularities in these systems is to compare and identify results that are outside the expected range. Such results may be the result of software errors, carelessness when entering data or, in the worst case, unethical behavior of students. The system should analyze and compare the results of all course participants, where deviations could indicate potentially suspicious activities. For effective recognition of such data and identification of suspicious cases, it is necessary to apply Data Mining techniques that will thoroughly analyze all available information.

IV. CONCLUSION

The purpose of education is to create expert engineers who can effectively respond to the demands of society. The education process should be constantly improved and supported, with quality courses and working conditions. Data Mining techniques can be applied from faculty selection to analyzing study program data to improve them. This data can be useful for planning the enrollment policy and improving the teaching process.

The application of Data Mining techniques in higher education opens up many opportunities for optimizing the learning process, supporting students and making informed decisions at the level of an educational institution. Analyzing data about students, their interactions with courses, grades, activities and other relevant factors enables a deeper understanding of their needs, behavior and performance.

One of the key advantages of Data Mining in higher education is the personalization of student support. Based on data analysis, the individual needs of students can be identified and appropriate resources and support can be adapted to them. This may include tailoring teaching materials, recommendations for additional courses or activities, and providing personalized advice for academic development.

Also, Data Mining makes it possible to predict the success of students. Analyzing past data on student performance can help identify factors that influence academic success and create models to predict future performance. Such models can be of great benefit both to students, providing them with insight into their opportunities and areas for improvement, and to educational institutions, allowing them to react in a timely manner and provide additional support to those who need it most.

In addition, Data Mining can contribute to the optimization of the teaching process. Analyzing data on student interactions with courses can help identify effective instructional strategies, identify gaps in course materials or practices, and adjust courses to better meet student needs.

Ultimately, the application of Data Mining techniques in higher education aims to improve the experience of students, increase the efficiency of the educational process and achieve better results. Integrating this technology into educational management strategies can help educational institutions use their resources more efficiently, provide better support to students and continuously improve the quality of education.

REFERENCES

[1]. Frawley, W.J., Piatetsky-Shapiro, G. andMatheus, C.J. [1992] "Knowledge Discovery in Databases: An Overview" AI Magazine, *13*(3), 57-70. DOI: https://doi.org/10.1609/aimag.v13i3.1011

[2]. Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. [1996] "From Data Mining to Knowledge Discovery in Databases" AI Magazine, *17*(3), 37-54. DOI: https://doi.org/10.1609/aimag.v17i3.1230

[3]. Batool, S., Rashid, J., Nisar, M.W., Kim, J., Kwon, H.Y. and Hussain, A. [2023] "Educational Data Mining to Predict Students' academic performance: A survey study" Education and Information Technologies, Vol. *28*, 905-971. https://link.springer.com/article/10.1007/s10639-022-11152-y

[4]. Sharma, A., Mansotra, V. and Mahajan, R. [2015] "Applying Data Mining in Higher Education Sector" International Journal of Software and Web Sciences (IJSWS), 88-92. https://www.researchgate.net/publication/350689978_Applying_Data_Mining_in_Higher_Education_Sector

[5]. Kačapor, K. and Lagumdžija, Z. [2020] "Rudarenje edukacijskih podataka: korištenje klasteriranja za predikciju studentskog uspjeha" Conference: 43. Međunarodni ICT skup MIPRO 2020, 1075-1080. https://www.researchgate.net/publication/351871935_Rudarenje_edukacijskih_podataka_koristenje_klasteriranja_za_predikciju_studentskog_uspjeha

[6]. Calders, T. and Custers, B. [1992] "What Is Data Mining and How Does It Work?" Discrimination and Privacy in the Information Society, Nr. 3 Heidelberg: Springer, 27-42. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3047758

[7]. Annas, M. and Wahab, S.N. [2023] "Data Mining Methods: K-Means Clustering Algorithms" International Journal of Cyber and IT Service Management (IJCITSM), *3*(1), 40-47. DOI: https://doi.org/10.34306/ijcitsm.v3i1.122

[8]. Manning, C., Raghavan, P. and Schütze, H. [2009] "An Introduction to Information Retrieval" Online edition, Cambridge University Press Cambridge, 350-400. https://www.academia.edu/27076940/An_Introduction_to_Information_Retrieval

[9]. Nisbet, R., Miner, G. and Yale, K. [2018] "Handbook of Statistical Analysis and Data Mining Applications" Second Edition, Academic Press, Elsevir. http://repo.darmajaya.ac.id/4157/1/Handbook%20of%20statistical%20analysis%20and%20data%20mining%20applications%20%28%20PDFDrive%20%29.pdf

[10]. Nava, J. and Hernández, P. [2012] "Optimization of a Hybrid Methodology (CRISP-DM) "DOI: 10.4018/978-1-4666-0297-7.ch014

[11]. CRISP-DM: Data Mining Process, This figure is attributed to Kenneth Jensen, Wikimedia Commons. https://app.myeducator.com/reader/web/1421a/2/qk5s5/

[12]. Qasrawi, R., Badrasawi, M., Al‐ Halawa, D.A., Polo, S.V., Khader, R.A., Al‐ Taweel, H., Alwafa, R.A., Zahdeh, R., Hahn, A. and Schuchardt, J.P. [2024] "Identification and Prediction of Association Patterns Between Nutrientintake and Anemia Using Machine Learning Techniques: Results From a Cross‐ sectional Study With University Female Students From Palestine" European Journal of Nutrition, Springer. https://doi.org/10.1007/s00394-024-03360-8

[13]. Goyal, M. and Vohra, R. [2012] "Applications of Data Mining in Higher Education" International Journal of Computer Science Issues, (IJCSI), Vol. 9, *2*(1), 113-120. https://www.researchgate.net/publication/264888758_Applications_of_Data_Mining_in_Higher_Education

[14]. Aji, I. and Sunyoto, A. [2020] "An Implementation of C4.5 Classification Algorithm to Analyze Student's Performance" 3rd International Conference on Information and Communications Technology (ICOIACT). DOI: 10.1109/ICOIACT50329.2020.9332088

[15]. Juma, S., Shaharanee, I.N.M. and Jamil, J.M. [2023] "Clustering Student Performance Data Using k-means Algorithms" Journal of Computational Innovation and Analytics (JCIA), *2*(1), 41-55. DOI: 10.32890/jcia2023.2.1.3

[16]. Chi, D. [2021] "Research on the Application of K-Means Clustering Algorithm in Student Achievement" International Conference on Consumer Electronics and Computer Engineering (ICCECE), 435-438. https://sci-hub.yt/10.1109/iccece51280.2021.9342164

[17]. Duhaim, A.M., Al-mamory, S.O. and Mahdi, M.S. [2022] "Cheating Detection in Online Exams during Covid-19 Pandemic Using Data Mining Techniques" Webology, (19)1. DOI: 10.14704/WEB/V19I1/WEB19026