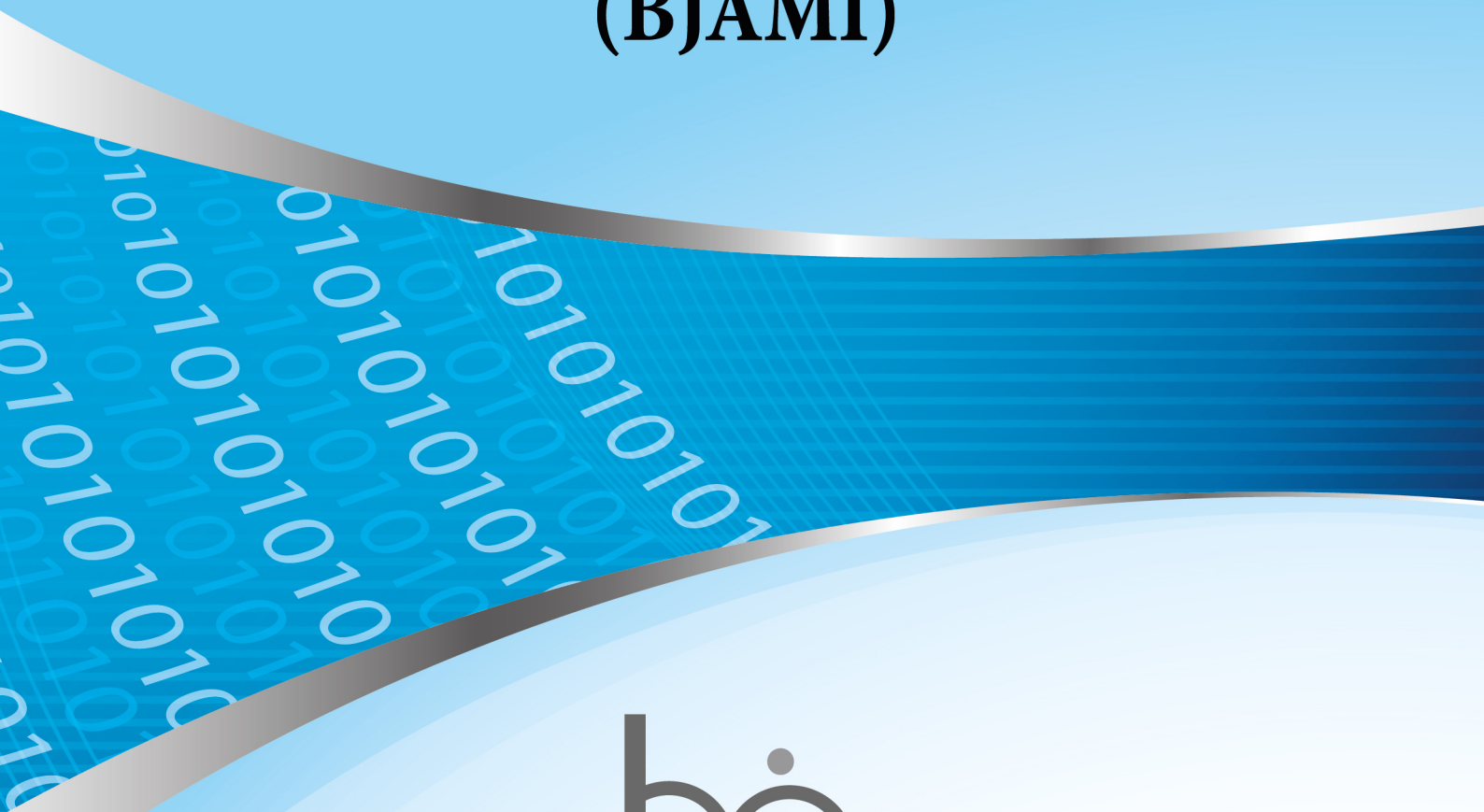
GOCE DELCEV UNIVERSITY - STIP
FACULTY OF COMPUTER SCIENCE

The journal is indexed in
EBSCO

ISSN 2545-4803 on line
DOI: 10.46763/BJAMI

BALKAN JOURNAL OF APPLIED MATHEMATICS AND INFORMATICS (BJAMI)



BALKAN JOURNAL
OF APPLIED MATHEMATICS AND INFORMATICS

YEAR 2022

VOLUME V, Number 2

AIMS AND SCOPE:
BJAMI publishes original research articles in the areas of applied mathematics and informatics.

Topics:
1. Computer science;
2. Computer and software engineering;
3. Information technology;
4. Computer security;
5. Electrical engineering;
6. Telecommunication;
7. Mathematics and its applications;
8. Articles of interdisciplinary of computer and information sciences with education, economics, environmental, health, and engineering.

Address of the editorial office
Goce Delcev University – Stip
Faculty of philology
Krste Misirkov 10-A
PO box 201, 2000 Štip,
Republic of North Macedonia

BALKAN JOURNAL
OF APPLIED MATHEMATICS AND INFORMATICS (BJAMI), Vol 5

ISSN 2545-4803 on line
Vol. 5, No. 2, Year 2022

EDITORIAL BOARD

CONTENT

Balkan Journal of Applied Mathematics and Informatics
Volume V Number 2 Year 2022
Online ISSN 2545-4803
UDC: 004.85:004.422.635.3

PREDICTING TEXTBOOK MEDIA SELECTION USING DECISION TREE ALGORITHMS

SADRI ALIJA[1], ALAA KHALAF HAMOUD[2] AND FISNIK MORINA[3]

Abstract. In this paper, we will investigate machine learning algorithms in predicting the medium of reading whether it be digital or paper based. The dataset are collected based on a survey that was anonymous and conducted among students at the South East European University (SEEU) in the Republic of North Macedonia (RNM) and the Faculty of Business of Haxhi Zeka University (HZU) in the Republic of Kosovo. 262 respondents are surveyed, with 180 female (68.7%) and 82 male (31.3%) students. The Synthetic Minority Oversampling Technique (SMOTE) filter is applied to oversample the dataset without affecting the original data and to find the optimal accuracy of algorithms. Four decision tree algorithms are examined (J48, Random Forest, Random Tree, and Rep Tree) over the data before and after applying SMOTE filter based on the performance criteria (TP rate, FP rate, Precision and Recall). Random Forest proved its accuracy with (0.743) for precision and (0.733) for recall over the other three algorithms.

Keywords. Decision Tree, SMOTE filter, Machine learning, Textbook selection

1. Introduction

With the increase in the amount of human knowledge, there was a need to write it down. In the beginning, these were manuscript books, which was a long and expensive process of creating books, and with the advent of printing, the invention of the printing press replaced the manuscript book with a printed book [1][2]. Today, we cannot say that e-books have replaced printed books, but they have certainly become more interesting and intriguing for new generations of users. Thus, by changing the paradigm, there was a need for a completely different form of organizing information and, ultimately, human knowledge. The explosion of information has opened new challenges, primarily regarding the storage of electronic records.

The answers presented in the paper should, at the very least, define the concept of an e-book, its users, and their habits, and also types of e-books, media, and formats. An e-book is a book in digital form, which is why we still call it a digital book or online book, and in English language also an e-book. Unlike traditional, classic, printed books, it represents a virtual, that is, screen structure, because the text is on the screen instead of on tangible, paper pages [3][4]. The most famous encyclopedia Britannica defines the term e-book as a digital file that contains text and images suitable for electronic distribution and display on the screen, in a way similar to a printed book. Many authors are trying to contribute to the definition of this term. According to the research of Vassiliou and Rowley (2008), 31 of the 37 included definitions start from the analogy with a printed book, meaning that an e-book is defined as a printed book. It is most often the idea that an e-book is an electronic version of a printed book, and that it has the characteristics of a printed book. Can the phenomenon of e-books be reduced to the phenomenon of new technology, or is this a new medium that necessarily brings with it new technology? These two aspects of the e-book, technological and media, go beyond the scope of this paper, but they are certainly crucial to mention. What matters is whether an e-book is used or read. From the technological aspect, e-books are used, and from the media aspect, e-books are used to be read [5].

The study process is a long-term process, and its ultimate goal is to make a person a built, self-conscious individual, capable of perceiving things and phenomena from the environment, critically analyzing them, and making decisions. Achieving this goal requires a lot of reading of required and recommended literature, but equally, if not more so, of informal literature of personal choice. It is very uncertain how much students read. This is difficult to measure and evaluate by some known methods. It is all, basically, up to the individual, who can freely decide how much and what he/she wants to read.

2. Literature Review

In [6], R. Junco, and C. Clem show that the outcomes of students learning are engaged with the digital textbooks usage. They used linear regression to predict the grades of students' courses according to their usage of digital textbooks. 233 students data are analyzed using linear regression to find if the metrics of digital textbook usage can predict the grades of the students' courses. They found that the variable of "time spent reading" strongly affected the prediction process. They also found that the top 10% of the highlights has higher grades in the final courses. The study found that the textbook analytics had great effect on the students' grade prediction.

Next, in [7], M. Kang et al investigated learner satisfaction and academic achievement when using digital textbooks based on the prediction process. They found that different variables affected user satisfaction and academic achievement such as learning attitude, teaching presence, cognitive presence, computer and academic self-efficacy, information literacy, direct-learning and prior knowledge of the student. Selecting the type of preference is investigated in [8], where 37 children are examined to find their preferences of reading digital or printed books. Parents are surveyed to investigate children's preference. The multimodal discourse analysis and descriptive statistics were used for analysis, where the study found that 27% of the predictions were accurately predicted. 65% of the choices related to children go to digital books. Reading digital books increases the physical position to the reader, discourse and attention of the children. 50% of children who read digital books request more books.

In [9] Yungwei Hao and Kathy Jackson examined students' satisfaction with using digital textbooks. 115 students in the undergraduate studies participated in this study to measure the satisfaction levels of using digital textbooks in two courses. The learning styles instrument was used to determine the learning preferences for students. The findings show that few factors emerged related to students' satisfaction such as learning facilitation and usability of digital textbooks.

In [10], Dmitriy V. Chulkov and Jason VanAlstine analyzed the student choices between digital and printed textbooks. The researchers did not find any relationship between the demographic characteristics and the student performance with the student choice of a textbook type. The medium selection is affected by the ability to keep and cost, while the learning style and ease of use are the factors that are considered by students. The multiple media of textbooks may allow students to work better.

The effect of reading media (digital or printed), and the time frame of reading (free or pressured time) on the attention of the readers is investigated in [11] by Pablo Delgado and Ladislao Salmerón. They found in their conclusions that in a pressure condition, on screen readers comprehended less than other groups of readers. While the outcomes of children's learning with printed or digital books are examined by Furenes, May Irene, Natalia Kucirkova, and Adriana G. Bus in [12], where they found that there are lower scores of comprehension for digital books. Printed books reading with the assistance of adults gave more effectiveness than reading digital books independently with enhancements. The enhanced digital books outperformed printed books in the context of story-congruent.

3. Model implementation

The implementation process of the model goes through five steps: data collection, data preprocessing, feature selection, implementing SMOTE, applying decision tree algorithms, and result evaluation, as shown in figure (1). The questionnaire implementing process is considered part of the data collection step. It involves preparing the dataset for the next steps, such as removing incomplete answers, improving the dataset's quality, converting columns' domains, and deriving the last column (the goal class). The goal class that drives the prediction process is the type of book: a printed book, an electronic book, or both. The dataset after data preprocessing goes through two directions, the first is the SMOTE filter and then applying feature selection to find the effect of the SMOTE filter on the dataset before/after the feature selection process.

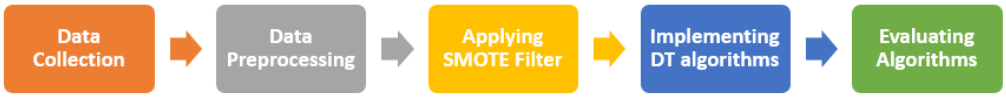


Figure 1. *Model Implementation Diagram*

3.1. Data Collection

The survey was anonymous and conducted among students at the South-East European University (SEEU) in the Republic of North Macedonia (RNM) and the Faculty of Business of Haxhi Zeka University (HZU) in the Republic of Kosovo. Of the 262 students surveyed, 180 were female (68.7%) and 82 were male (31.3%). Therefore, the survey was mostly comprised of female students, which will show the results of the survey to a greater extent and more accurately through the attitudes of the female part of the student generation belonging to generation Z towards books. In the survey, we conducted a baseline study examining students' preferences related to e-books and printed books in the context of reading materials for their study.

The survey is designed to collect these types of factors, demographic information about students, preference for reading books, agreement and disagreement with the feedback items for e-books (Satisfaction, Usefulness and Behavioral intention).

The Question possible value was shortened and converted from a nominal to a numeric type for ease of use and understanding. Response values of questions (Q13-Q28) are of the form {1; 2; 3; 4; 5} where 1; 2; 3; 4; 5 represents the answers "Strongly Disagree'', "Disagree", "Neutral", "Agree"," Strongly Agree" respectively. The questions related to each factor are described and listed in Table 1.

Table 1. *Survey Structure*

Seq	Question	Answer
Q1	Your gender	1=Male, 2=Female
Q2	University	1=SEEU, 2=UHZ
Q3	Year of Study	1=First, 2=Second, 3=Third, 4=Fourth
Q4	Experience using a computer/laptop	1=1-3 Years, 2=3-5years, 3=more than 5 years
Q5	Number of computers / laptops in the home	1=1, 2=2, 3=3 or more
Q6	The number of e-mail accounts	1=1, 2=2, 3=3 or more
Q7	Frequency of e-mail use	1=Every day, 2=Once a week, 3=Once a month, 4=Never
Q8	Frequency of text message use by mobile phone	1=Every day, 2=Once a week, 3=Once a month, 4=Never
Q9	Reading e-books	1=Yes, 2=No
Q10	Often reading e-books	1=Rarely, 2=Sometimes, 3=Always
Q11	Reading e-books During the	1=More than before the pandemic, Less

	pandemic Covid-19	than before the pandemic, 3= Same before and during the pandemic
Q12	Book format preference for academic studies	1=Printed books, 2=Electronic books, 3=Both printed and electronic books
Q13	I prefer using electronic books for my information needs over printed books	1=Strongly disagree, 2=disagree, 3=Neutral, 4=agree, 5=strongly agree
Q14	I consider electronic books capable of providing adequate information to meet my study needs	1=Strongly disagree, 2=disagree, 3=Neutral, 4=agree, 5=strongly agree
Q15	I am satisfied with the functions of e-books	1=Strongly disagree, 2=disagree, 3=Neutral, 4=agree, 5=strongly agree
Q16	I am satisfied with using an e-book as a learning assisted tool	1=Strongly disagree, 2=disagree, 3=Neutral, 4=agree, 5=strongly agree
Q17	I am satisfied with using an e-book as a reading tool	1=Strongly disagree, 2=disagree, 3=Neutral, 4=agree, 5=strongly agree
Q18	I am satisfied with the colors used in the e-book	1=Strongly disagree, 2=disagree, 3=Neutral, 4=agree, 5=strongly agree
Q19	The font size and typeface in the e-book were easy to read	1=Strongly disagree, 2=disagree, 3=Neutral, 4=agree, 5=strongly agree
Q20	The sequence of the e-book on the screen was clear	1=Strongly disagree, 2=disagree, 3=Neutral, 4=agree, 5=strongly agree
Q21	It was convenient to scroll up/down within the e-book	1=Strongly disagree, 2=disagree, 3=Neutral, 4=agree, 5=strongly agree
Q22	It was easy to turn the pages in the e-book	1=Strongly disagree, 2=disagree, 3=Neutral, 4=agree, 5=strongly agree
Q23	It was easy to find important information in the e-book	1=Strongly disagree, 2=disagree, 3=Neutral, 4=agree, 5=strongly agree
Q24	I intend to increase my use of e-books	1=Strongly disagree, 2=disagree, 3=Neutral, 4=agree, 5=strongly agree
Q25	I intend to use e-books to assist my learning	1=Strongly disagree, 2=disagree, 3=Neutral, 4=agree, 5=strongly agree
Q26	I will read e-books to look for the information I need	1=Strongly disagree, 2=disagree, 3=Neutral, 4=agree, 5=strongly agree
Q27	In five years, I will do most of my reading from e-books	1=Strongly disagree, 2=disagree, 3=Neutral, 4=agree, 5=strongly agree
Q28	If given a choice between an electronic or printed version of a particular book, I will choose the electronic version	1=Strongly disagree, 2=disagree, 3=Neutral, 4=agree, 5=strongly agree

3.2. Data Preprocessing

The dataset conducted from the questionnaire consists of 262 responses. In the data preprocessing stage, the records with empty answers on any question were removed to ensure the data consistency. The remaining responses were 234. This sample data is considered excellent with 0.932 Cronbach's alpha as shown in Table 2. The visualization process for the dataset is performed by Weka 3.8.5 tool, where the same tool is used to perform the next three stages of model implementation. For data reliability and consistency, the Cronbach's alpha value 0.932 is considered reliable for the 234 respondents and 28 questions [13][14], [15].

Table 2. *Cronbach's Alpha Value*

Number of respondents	Respondent Percentage	Number of Items	Cronbach's alpha
234	100%	28	0.923

3.3. Applying SMOTE Filter

The dataset is said to be imbalanced if the final class categories are not equally represented. The final class categories in real data sets are composed of a normal percentage of interesting examples and a small percentage of abnormal examples, with some abnormal examples misclassified as normal. Many applications find that this situation is critical and needs to be handled. For example, in fraud detection, if the data set holds 100.000 cases with normal examples and 1 abnormal case, there is a need to handle this case and avoid it. The problem of imbalanced data was faced in many attempts in different sectors such as text classification, telecommunication management, and fraudulent telephone calls. In the Synthetic Minority Over-Sampling Technique (SMOTE), the minority class (the low represented class in the final class) is oversampled by creating new synthetic ones. The extra training data is created based on the real data by using feature space rather than data space. The new synthetic examples are introduced along the *k* nearest neighbor's minority class.

The *k* nearest neighbors then are selected randomly according to the amount of the required over-sampling data [16]. The final class of the questionnaire data set consists of three classes (Electronic, Printed, and Both). The class distribution of the final class is for the Electronic (27), Printed (114), and Both (93). The SMOTE filter is applied to balance the final class categories in order to examine the decision tree algorithms and predict the type of reference. The final class distribution after two iterations of applying the SMOTE filter makes for Electronic (108), Printed (114), and Both (93).

3.4. Applying DT algorithms

The tool used for applying SMOTE filter, DT algorithms, and evaluating the performance is the Weka 3.8.5 tool. The stages of implementing DT algorithms (J48, Random Forest, Random Tree, and Rep Tree) are performed with the dataset before and after applying the SMOTE filter. The decision trees build a graph similar to the flowchart tree structure. The nodes are either non-leaf of leaf where the non-leaf ones are the test on specific variable, while the leaf ones are the results of the test. The final leaf in the tree represents the final class in the data set where the classification processes are implemented to find the prediction result.

The data set with numeric and categorical data types can be processed with different kinds of DT algorithms where the tree building process depends on the best variable that divides the classes. J48 DT algorithm is the extension of ID3 algorithm where it involves many features like pruning DT, rules derivation, dealing with continuous attribute values and missing values. Random Forest is the combination of predictors where each tree predictor depends on an independently sampled values of a random vector with the same value distribution for all forests' trees [17].

Rep Tree is one of DT algorithms that used the regression tree and created multiple trees in many different iterations and selected the best tree among all the trees that were created before. Mean Square Error is used for the pruning process based on the predictions of the tree [18]. Random Tree classifier perfectly fits the training data. The base learner of random tree is constructed and all data points are places in the root node. The non-terminal nodes are split randomly and examined until each two data points belong to two different classes [19][20].

3.5. Results Evaluation

The final stage includes evaluating the performance of DT algorithms according to performance criteria (True Positive rate (TP), False Positive rate (FP), precision and recall). The model implementation is performed according to the resampling method of (10-fold cross validation), to measure the accuracy of the model after resampling (train, and test) datasets. The confusion matrix in Table 3 represents the real and predicted classes. TP rate represents the positive cases which are predicted as positive, while FP rate represents the negative cases that are predicted as positive. True Negatives (TN) rate represents the negative cases that are predicted as negative while False Negative (FN) rate represents the negative cases that are predicted as positive. The precision value represents the ratio of predicted positive cases while the recall represents the ratio of relevant values that are detected.

Table 3. *Confusion Matrix*

Real	Predicate	
	Positive class	Negative class
Positive class	True Positive (TP)	False Negative (FN)
Negative class	False Positive (FP)	True Negative (TN)

The performance criteria are listed for each factor before and after applying the SMOTE filter. The TP rates are enhanced for all algorithms where the best performance goes to the Random Forest algorithm with 0.733 after applying the SMOTE filter when its value is 0.650 before applying the filter. Random Tree TP rate is 0.556 before applying the filter and 0.625 after applying the filter which is considered as the lowest TP rate among all algorithms.

Random Forest FP rate is the lowest among all algorithms with 0.266 before applying the filter and 0.132 after applying the filter, while J48 FP rate is 0.304 before applying the filter and 0.182 after applying the filter, which is considered the highest among all other algorithms. Precision value for Random Forest considers the best value is 0.656 before applying the filter, and 0.743 after applying the filter, while the lowest precision value is 0.569 before applying the filter and 0.630 after applying the filter for J48. Finally, the recall value of Random Forest is the best with 0.650 before applying the filter and 0.733 after applying the filter, while the lowest value is 0.556 before applying the filter and 0.625 after applying the filter for Random Tree. Table 4 shows the performance criteria of DT algorithms before and after applying the SMOTE filter.

Table 4. *Performance Results of Decision Tree Algorithms before/after Applying the Smote Filter*

Algorithm		J48	Random Forest	Random Tree	Rep Tree
TP rate	Before	0.573	0.65	0.556	0.53
	After	0.629	**0.733**	0.625	0.648
FP rate	Before	0.304	0.266	0.287	0.345
	After	0.182	**0.132**	0.18	0.172
Precision	Before	0.569	0.656	0.564	0.521
	After	0.63	**0.743**	0.638	0.651
Recall	Before	0.573	0.65	0.556	0.53
	After	0.629	**0.733**	0.625	0.648

The optimal algorithm in predicting the type of reference according to Table 4 is Random Forest. SMOTE filter has the effect of the resulting performance criteria where accuracy is enhanced after its application. The effect shows that the size of the data set, besides the distribution of final class categories, can affect the accuracy of the algorithm in predicting the final class.

4. Conclusion, discussion, and future works

Most of the research, which focused on students' reading habits, actually started from the fact that psychological, social, and financial conditions significantly influence the reader, modify and re-create his/her reading habits. Thus, depending on the current state or status of the reader, a person can read more or less, consume more demanding or easier reading, understand more complex topics and texts. Therefore, research can be based on the difference between reading for educational and leisure purposes. Education requires continuous consumption of several different texts, articles, and books, while entertainment for the student population is often not reading, but some other social activities such as going out, traveling, sports, etc. The goal was to determine the reasons for reading, what kind of literature students read, and on what medium. The largest percentage had the offered answer that the motive for reading is personal development, and the choice of literature is based on the genre/topic they prefer or need at that moment. When it comes to the type of literature they read, most students go for newspapers and magazines, followed by professional literature, literature, and the portals they follow.

Selecting the printed books or digital books may differ from student to student. This study investigates the prediction process of textbook media using four decision tree algorithms (Random Forest, Rep Tree, Random Tree, and J48). The results show that the random forest algorithm can predict the textbook media more accurately than other decision tree algorithms (J48, Random Tree, and Rep Tree). Random Forest decision tree after applying the SMOTE filter outperforms Random Tree, J48, and Rep Tree decision trees in both TP rate and recall with (0.733). The algorithm also outperforms the other algorithms in precision with (0.743), and in FP rate with (0.132). In future works, the model will be developed to predict the factors influencing the selection of textbooks media in order to find the effect on students' performance. The factors such as time spent on reading, geographical characteristics, health, and teaching presence on textbook media selection had the direct effect on the student performance.

References

[1] E. Buringh and J. L. Van Zanden. (2009), Charting the 'Rise of the West': Manuscripts and Printed Books in Europe, a long-term Perspective from the Sixth through Eighteenth Centuries, *J. Econ. Hist.*, vol. 69, no. 2, pp. 409–445.

[2] I. Horvat. (2016), Kultura čitanja u 21. stoljeću, University North. University centre Koprivnica. Department of Journalism.

[3] G. A. Strouse and P. A. Ganea, (2017). A print book preference: Caregivers report higher child enjoyment and more adult--child interactions when reading print than electronic books, *Int. J. Child-Computer Interact.*, vol. 12, pp. 8–15.

[4] T. Morineau, C. Blanche, L. Tobin, and N. Guéguen, (2005). The emergence of the contextual role of the e-book in cognitive processes through an ecological and functional analysis, *Int. J. Hum. Comput. Stud.*, vol. 62, no. 3, pp. 329–348.

[5] B. Badurina and H. Serdarušić, (2015). Percepcija e-knjige studenata u Hrvatskoj, *Libellarium časopis za istraživanja u Pod. Inf. i Srod. Znan.*, vol. 8, no. 2, pp. 65–80.

[6] R. Junco and C. Clem, (2015). Predicting course outcomes with digital textbook usage data," *Internet High. Educ.*, vol. 27, pp. 54–63.

[7] M. Kang, M. Kim, S. Yoon, and H. Lim, (2011) Investigating the Predicting Variables of the Learning Outcomes when using Digital Textbooks, in *EdMedia+ Innovate Learning*, 2011, pp. 160–169.

[8] L. Eutsler and J. Trotter, (2020). Print or iPad? Young Children's Text Type Shared Reading Preference and Behaviors in Comparison to Parent Predictions and At-home Practices, *Lit. Res. Instr.*, vol. 59, no. 4, pp. 324–345.

[9] Y. Hao and K. Jackson, (2014). Student satisfaction toward e-textbooks in higher education, *J. Sci. Technol. Policy Manag*.

[10] D. V Chulkov and J. VanAlstine, (2013). College student choice among electronic and printed textbook options, *J. Educ. Bus.*, vol. 88, no. 4, pp. 216–222, 2013.

[11] Delgado, Pablo, and Ladislao Salmerón. "The inattentive on-screen reading: Reading medium affects attention and reading comprehension under time pressure." Learning and instruction 71 (2021): 101396.

[12] Furenes, May Irene, Natalia Kucirkova, and Adriana G. Bus. "A comparison of children's reading on paper versus screen: A meta-analysis." Review of educational research 91.4 (2021): 483-517.

[13] A. Lazar, (2019). Simbolizam knjige u generaciji Z, Josip Juraj Strossmayer University of Osijek. Faculty of Humanities and~….

[14] G. D. Israel, (1992). Determining sample size.

[15] U. Sekaran and R. Bougie, (2016). *Research methods for business: A skill building approach*. john wiley & sons.

[16] U. Sekaran and R. Bougie, (2003). Research methodology for business. New York: John Wiley & Sons, Inc.

[17] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, (2002). SMOTE: synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357.

[18] L. Breiman, "Random Forests, (2001). *Mach. Learn.*, vol. 45, no. 1, pp. 5–32.

[19] S. K. Jayanthi and S. Sasikala, (2013). eptree classifier for identifying link spam in web search engines, *IJSC*, vol. 3, no. 2, pp. 498–505.

[20] A. Cutler and G. Zhao, (2001). Pert-perfect random tree ensembles, *Comput. Sci. Stat.*, vol. 33, pp. 490–497.

SADRI ALIJA
Faculty of Business and Economics
SouthEast European University, Tetovo, North Macedonia
E-mail address: s.aliji@seeu.edu.mk

ALAA KHALAF HAMOUD
University of Basrah, Basrah, Iraq
E-mail address: alaa.hamoud@uobasrah.edu.iq

FISNIK MORINA
Faculty of Business
Haxhi Zeka University, Peja, Kosovo
E-mail address: fisnik.morina@unhz.eu