Predictive Modeling of Student Performance in Moodle LMS using Learning Analytics

Mužinić Mia, Antonela Sikavica, Petra Zelić, Ani Grubišić, Ines Šarić-Grgić
mmuzinic@pmfst.hr, asikavica@pmfst.hr, pzelic@pmfst.hr, ani@pmfst.hr, insaric@pmfst.hr

University of Split, Faculty of Science Split, Croatia

Abstract—**This research paper explores the utilization of learning analytics to predict students' performance in Moodle Learning Management System (LMS). This study aims to analyze the accuracy of various predictive models in determining student success in passing or failing a course and predicting their final grades. The performance of ten different models was evaluated using a dataset of students records within Moodle, including demographic information, engagement patterns, and academic outcomes. The results revealed significant variations in accuracy across the models. The Gradient Boosting Classifier achieved accuracy of 96.296% in predicting course outcomes, while the Random Forest model demonstrated accuracy of 92.593%. In terms of predicting final grades, the Random Forest Classifier performed with an accuracy of 92.593%, and the Decision Tree Classifier and Gradient Boosting Classifier also achieved with 92.593%, with other models also exhibiting reasonably good accuracy. These findings provide valuable insights into the potential of predictive modeling techniques for teachers and students seeking to identify students in danger of not succeeding and enhance their academic outcomes. Further research is recommended to explore the underlying factors and additional features among test scores that can contribute to predictive modelling contributing to gender-based variations in student success and to refine the models for improved performance.**

Index terms—**Moodle LMS, e-learning systems, educational data mining, predicting final grades, student performance, student success**

I. Introduction

Ever since personal computers and the Internet became widespread, numerous aspects of life, including education, have witnessed remarkable transformations. Higher education institutions have enthusiastically embraced e-learning platforms like Edmodo, Canvas, Schoology, Blackboard Learn, Sakai, Moodle, ATutor, and Chisimba. As a consequence, e-learning has emerged as a subject of extensive study, with numerous top-tier publications endorsing the use of digital learning approaches. The importance of e-learning technology has consistently increased, causing many educational institutions to view it as an essential component of their academic curricula. The growing integration of Learning Management Systems (LMS) in educational institutions has opened up extensive possibilities for the accumulation and assessment of learning data. Learning analytics, a field that leverages this data to gain insights into students' learning behaviors and results, offers significant potential for enhancing educational approaches.

In this study, we explore the use of learning analytics methods for forecasting student performance in Moodle, a widely used LMS platform. At the heart of a Learning Management System (LMS) is the goal of structuring courses and efficiently administering learners in an online learning environment. The advancement of online learning systems assumes a critical role in promoting vital '21st-century skills' such as critical thinking, creativity, and communication, which are central to students' achievements in higher education and future careers. E-learning presents notable advantages over traditional classroom learning, chiefly in terms of accessibility and flexibility. The availability of online resources empowers individuals to access information effortlessly, facilitating learning anytime and anywhere. In the beginning, Learning Management Systems (LMSs) were fairly basic, resembling simple web pages that provided information about different disciplines, lecturers, and teaching methods. However, over time, they have evolved into complex and versatile platforms. Modern LMSs not only facilitate the exchange of educational materials but also enable interactive communication between teachers and students. Furthermore, 1 Learning Management Systems (LMSs) enable impartial evaluation and the statistical scrutiny of students' mastery of course materials. Contemporary e-learning systems make use of advanced technologies and features that support various educational activities, including face-to face interactions. Some programs even offer individuals the opportunity to become students and earn a university diploma without having to leave their homes. LMSs are employed for the formation of study groups, the delivery of lectures, seminars, practical sessions, and the management of examinations.

Within the context of forecasting student achievements in the Moodle Learning Management System (LMS) through the utilization of Learning Analytics, the following research questions are the focus of this researching:

1. How effective are learning analytics in predicting students' performance in Moodle LMS?
2. How accurate are the predictive models built using learning analytics in Moodle LMS?

3. How do different factors, including factors like gender or previous academic performance, influence the effectiveness of predictive models in Moodle LMS?
4. Can learning analytics in Moodle LMS be used to personalize and tailor instructional approaches to individual students' needs?

The main contributions of this paper are valuable understandings into the effective application of learning analytics to forecast students' achievements within the Moodle Learning Management System (LMS), a comprehensive summary of the significant advancements made by famous researchers in this particular field, highlighting a key challenges and open research questions in predicting students' performance.

The second section delves into studies on predicting student success through learning analytics, offering insights into crucial contributions, methodologies, and findings from reputable sources. In the third section describe the methodology used in our research. In the Results section, we analyzed various factors influencing student success and developed predictive models for course completion and final grades using Python and sklearn. Section Discussion assesses how well learning analytics predict student success in Moodle LMS. Also, in this section we explore the link between gender and academic performance and consider the potential use of learning analytics for personalized teaching. The Conclusion summarizes steps for predicting student data, discusses challenges and future plans, and emphasizes the potential benefits for teachers and students using Moodle data.

II. RELATED WORK

To conduct a thorough literature review, it is essential to start by selecting relevant studies that directly address the research objectives. This involves a meticulous process of identifying papers that align with the research concerns and meet specific inclusion and exclusion criteria. By carefully defining these criteria, the literature review maintains focus and ensures the inclusion of high-quality studies. This section provides an overview of previous research in predicting student educational achievements, both in a broader context and specifically in the realm of programming education. The rise of online learning environments, especially during the recent pandemic, has led to an increase in their usage and the generation of vast amounts of data. This data provides researchers with valuable opportunities to conduct more relevant and impactful studies using learning analytics.

To identify significant contributions, an extensive literature search was performed using reputable databases such as the Social Sciences Citation Index (SSCI), the Science Citation Index Expanded (SCIE), and the IEEE Xplore database. The exploration concentrated on vital fields including titles, abstracts, and keywords of publications listed in the Web of Science Core Collection and IEEE Xplore. Specific search keywords were used, combining different fragments with the AND operator. The initial segment included terms associated with the learning analytics and EDM, so the term (Learning Analytics* OR Educational data mining*) was employed as the opening portion of the search query. To encompass e-learning systems, the second section of the search query was constructed in the following manner: (Moodle LMS* OR E-learning systems*). To avoid all papers that do not contain students success, a third fragment (Student performance* OR Student success*) is included. Ultimately, the distinct query was formulated by combining the first three segments, as elucidated at the start of this section, and linking them with an AND operator: (Learning Analytics* OR Educational data mining*) AND (Moodle LMS* OR E-learning systems*) AND (Student performance* OR Student success*). This comprehensive approach aimed to capture relevant publications and ensure the inclusion of influential works in the review process. We also utilized Google Scholar, an online search engine for scientific literature, to find relevant articles on the topics of Learning Analytics, Moodle LMS, elearning systems, educational data mining, predicting final grades, student performance, student success. We conducted the search using specific keywords and selected articles based on their titles, abstracts, and the number of citations they received.

We have identified key researchers in this particular field and made a comprehensive summary of their significant advancements. One of them is Ryan Baker, a distinguished scholar in the domain of learning analytics, recognized for his contributions in educational data mining, learning analytics, and the utilization of data-driven strategies to enhance educational achievements. Baker has conducted extensive research on topics such as student engagement, intelligent tutoring systems, and educational technology. His research frequently entails the examination of extensive educational data to acquire understanding regarding student conduct, learning methodologies, and the efficiency of instructional methods. Baker's research has contributed to the advancement of the field and has been influential in shaping the use of learning analytics in education [2, 3, 11]. George Siemens is known for his work on connectivism and the application of learning analytics in understanding and improving digital learning environments [10, 11]. Additionally, Shane Dawson is well-known for his contributions in the realm of learning analytics, educational data mining, and the utilization of data-driven techniques to enrich teaching and learning results [7] and Dragan Gašević who is acknowledged for the research in learning analytics, EDA, and the development of efficient learning analytics methods and tools [6].

The educational scientific community has been grappling with the challenge of predicting students' academic performance. In this paper, we review the existing literature that aligns closely with our approach. Specifically, we focus on studies that meet the following criteria: they predict students' performance based on information gathered from Learning Management Systems (LMS). In addition to the Zhang et al. (2020) [13] article, we have also found some other relevant articles related to this field.

Romero et al. (2013) [9] sought to forecast students' ultimate grades by leveraging Learning Management System (LMS) data, encompassing details about both successfully and unsuccessfully completed quizzes and assignments, in addition to the time invested in quizzes, forums, and assignments. Their study entailed assessing the efficiency of various data mining techniques for encompassing statistical methods, classifying students, rule induction, neural networks etc. The analysis was conducted across seven different Moodle courses. The results indicated that the CART (Classification and Regression Trees) decision tree algorithm achieved the best performance among the classification methods, with an accuracy rate of 65%. This suggests that the CART decision tree model outperformed the other techniques in predicting students' final grades using the available LMS data.

In their study, Gašević et al. (2016) [6] developed logistic regression models to predict students' outcomes in nine undergraduate courses. Two types of models were built: one encompassing all the courses together, and individual models for each course. The performance of the models was assessed using the area under the ROC (Receiving Operating Characteristic) curve values. The model that considered all the courses achieved an acceptable level of accuracy, with an AUC value between 0.5 and 0.7. Nonetheless, the models tailored for each individual course exhibited outstanding performance (AUC between 0.8 and 0.9) or outstanding performance (AUC greater than 0.9). This indicates that while the model considering all the courses provided a reasonable level of accuracy, the models tailored to specific courses exhibited notably higher predictive performance. These findings highlight the importance of developing course specific models for predicting students' performance, as they yield more accurate and reliable results.

Conijn et al. (2017) [4] conducted an analysis of hybrid courses and students utilizing the Moodle Learning Management System (LMS). Their primary goal was to forecast students' ultimate grades by utilizing LMS predictor variables alongside interim assessment grades. They utilized logistic regression models for pass/fail forecasts and standard regression models for the prediction of final grades. The prediction was made for the first 10 weeks of the course. The accuracy of the predictions showed a slight improvement as the prediction week progressed. For instance, by week 5, the regression model achieved an adjusted accuracy of 43%, whereas the binary classifier for pass/fail prediction reached a 67% accuracy by week 3.

In contrast to their system, where Conijn et al. (2017) [4] didn't develop particular models for detecting students with outstanding performance or those at severe risk, Riestra-Gonzalez et al. (2021) [8] employed ML 3 techniques to develop models for the early anticipation of students' success in addressing LMS assignments, where their analysis focuses solely on the LMS log files created up to the prediction time. Their objective is to forecast students' outcomes at specific milestones within the course, including 10%, 25%, 33%, and 50% of its duration. Their objective was not to predict the exact marks students would receive on LMS assignments. Instead, their effort was directed towards recognizing students who were at risk, prone to failure, or exhibited exceptional performance in the initial phases of the course. By emphasizing early identification, their models offered valuable insights that could enable timely interventions and support for students in need of additional assistance or those seeking more challenging coursework.

Costa et al. (2017) [5] conducted a comparison of the efficacy of various EDM techniques in identifying students who are prone to struggling in two introductory programming courses. The initial course was administered as a distance learning program, encompassing 262 students across a span of 10 weeks. In contrast, the second course was conducted in an on-campus setting and comprised 161 students. Unlike the system described in this article, the datasets employed by Costa et al. encompassed not only LMS interaction data but also additional variables like age, gender, marital status, location, income, enrollment year, field of study, and student performance in weekly assignments and exams. The most noteworthy evaluation metrics, such as accuracy and precision, for the first week, were reported as 0.77 and 0.8 for the two courses, respectively. These values improved as the prediction timeframe advanced. However, it is worth noting that the study did not focus specifically on a single course, and the analysis of only two courses may limit the creation of course-agnostic models.

Tomasevic et al. (2020) [12] conducted a study where they developed classification and regression models to predict student exam marks.Their study used the OULAD, which includes information on performance in course assessments, student demographics, and student engagement that goes beyond LMS logs alone. In their study, Tomasevic et al. focused on a classification model that distinguishes between students who pass and fail. Predictions were generated at different junctures, which encompassed just before the final exam and following several intermediate evaluations. The evaluation metrics showed an upward trend as the prediction point advanced. The metrics ranged from 78% to 94.9%. Regarding the forecast conducted right before the final exam, the evaluation metric reached 96.6%. Regression models were also developed to predict final exam scores, and the results followed a similar pattern. Among the various models used, in the research carried out by Tomasevic et al., it was determined that neural network models displayed the most impressive results for both regression and classification assignments.

III. Methodology

This study specifically focuses on the LMS Moodle. In this study, we referenced a paper titled "Using Learning Analytics to Predict Students' Performance in Moodle LMS" by Zhang et al. (2020) [13]. This study seeks to explore the efficiency of learning analytics in predicting students' performance in Moodle LMS by analyzing various data sources available within the system. We examine the relationships between students' demographic information, their engagement patterns, and academic outcomes to develop predictive models. Our goals for this research are to analyse the data obtained through the LMS Moodle and predict the final grade of students in the "Programming 1" course. However, there is a certain difference between the dataset used in the research paper by Zhang et al. (2020) [13] and the dataset we used in our research, which resulted in certain variations in the data analysis. The authors of the paper used data on the quantity of messages dispatched and perused on the forum for each student, while such records were not available in the dataset we used for our research. Additionally, in the referenced paper, the authors utilized data regarding the overall time students invested in quizzes, assignments, and the forum, which was assessed through correlation analysis. Furthermore, in our research, in addition to correlation analysis, we also employed predictive modeling, where, as mentioned before, we predicted whether a student passed or failed based on the tests they solved during the semester. Additionally, we analyzed the results of quizzes, as well as the number of passed and failed tests, to predict the student's final grade. By combining these factors, we developed predictive models that showed promising results in predicting student outcomes and grades. During our research, we followed the steps representing the general data lifecycle in data mining, as illustrated in Fig 1. Furthermore, a more detailed description of the predictive models is provided later in the article.

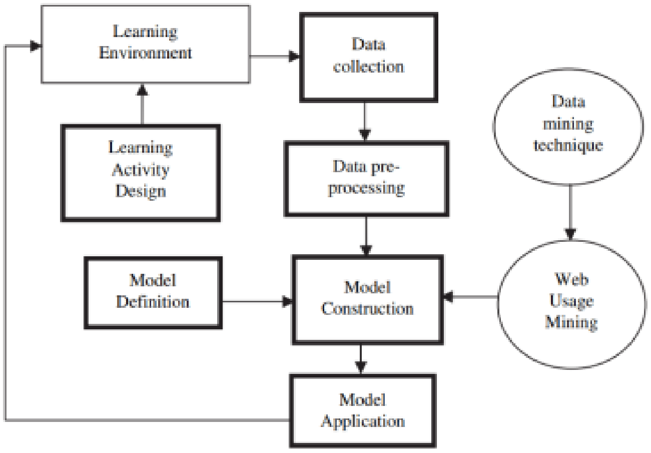

Fig. 1. Proposed Architecture framework for data-mining and E-learning, taken from the research paper [1]

LMS components provide opportunities for improving student learning outcomes and impacting final grades. The Moodle log captures valuable data such as access timestamps, IP addresses, student names, actions performed, module activities, and additional details. This data can be utilized for data mining, enabling research, visualization, and analysis to identify meaningful patterns in analyzing student conduct and feedback. Data mining continues to be a promising domain for investigating educational data. Customized learning analysis systems, including business intelligence tools and applications combining database analysis and student identification, are being developed. To enhance prediction of student success, clustering rules from the Moodle LMS module can be applied for data mining. This approach utilizes data on students' quantitative, qualitative, and social activities to improve the accuracy and interpretability of predictions. This study predicts whether a student passed or failed, and the student's ultimate grade determined by the use of learning materials. The research relied on the Moodle database as the data origin, and the data was formatted in CSV. In line with the research's specific requirements, the data collected included percentage scores from midterm exams and online tests.

Data preprocessing was carried out using Python libraries like NumPy and Pandas, as well as Excel for various data processing activities. The following information was extracted from a university student's Moodle log: Student ID, results of reviewed lessons and solved tests, total number of passed and failed tests, midterm exam1, midterm exam2, gender, is the student repeating the course or not, and the final grade. Our teacher Ani Grubišić has prepared the created data. Several machine learning algorithms, such as: Naive Bayes Algorithms, Logistic Regression, Gradient Boosting Classifier, Support Vector Machine, Stochastic Gradient Descent, Decision Tree Algorithm, Random Forest Classifier Algorithm, K-Nearest Neighbor Algorithm, and Neural Networks will be applied to develop predictive models. Feature engineering techniques will be employed to extract relevant features from the dataset.

The outcomes of this research will offer insights into the feasibility and accuracy of utilizing learning analytics for predicting student grades in Moodle LMS. By identifying the key predictors of academic success, educational establishments have the opportunity to proactively respond, enabling at-risk students to receive timely, tailored support, which positively impacts retention rates and overall learning outcomes.

A. Participants

Out of the total of 169 participants in the research conducted in the Programming 1 course at the Split's University Faculty of Science, 43% were male and 57% were female. We also found that 28% of the participants were repeaters and had attended the course before.

B. Study Design and Procedure

Throughout the semester, students were observed and information were gathered about their activity on a Moodle LMS. Some of the observed features are: lesson completion percentage, results on tests, Student ID, total number of passed and failed tests, midterm exams, gender, and final grade per student.

C. Data Collection and Processing

During the Introduction to Programming 1 course, Moodle served as a platform for providing students teaching materials, lessons, forums, as well as conducting tests. Log files about the student's activity on the platform contained valuable information such as the time spent on the platform, test completion percentage, gender, number of passed and failed tests and final grades. To conduct this research, a comprehensive dataset consisting of students' demographic data, interaction logs, assignment grades, forum participation, and quiz scores will be collected from a representative sample of Moodle users.roughout the semester, students were observed and information were gathered about their activity on a Moodle LMS. Some of the observed features are: lesson completion percentage, results on tests, Student ID, total number of passed and failed tests, midterm exams, gender, and final grade per student.

To facilitate further analysis, we performed data preprocessing on the collected data. In the context of data analysis, it is essential to examine the significance of variables and eliminate those that are superfluous or noisy. Consequently, we eliminated variables unrelated to the objectives of our study. The next stage includes dealing with unexplained or missing data, which often arise when students are not engaged in prescribed learning activities It is important to control for missing values during the data preprocessing phase for predictions model development. To address this, we used imputation techniques to estimate the missing values based on the training data. The choice of imputation method was based on obtaining the most effective measure of central tendency rather than missing values. Where students missed some courses or tests, we used zero imputation, as well as average imputation.

To address the variations in data distributions and ensure that our algorithms effectively capture the underlying patterns, we employed a normalization technique. Specifically, we utilized the MinMaxScaler estimator from the sklearn.preprocessing package. This process rescales the data, conveniently mapping the features to a range between 0 and 1. Additionally, we harnessed a statistical metric known as the correlation matrix, which calculates the Pearson coefficient to gauge the strength of linear relationships between different variables. In our research, the correlation matrix is used for statistically describing the data and aiding in feature selection. This selection process involves choosing input variables that make the most significant contributions to the prediction models.

D. Model Development

For the predictive data analysis, we employed the sklearn library, which is a simple and efficient tool for statistical and ML modeling that includes also classification. We developed model for predicting the pass/fail scores in practical exams for a particular course module, considering the online lessons and tests. Also, we developed a new model specifically for predicting students final grade, considering midterm exams and the number of passed and failed tests.

For the classification task, we employed several algorithms from the sklearn library: GaussianNB, BernoulliNB, Logistic Regression, Random Forest Classifier, Support Vector Machine, Decision Tree Classifier, KNeighbors Classifier, Gradient Boosting Classifier, Stochastic Gradient Descent, Neural Networks. Each model was trained using 80% of the gathered data and other 20% was used for testing. Ultimately, we assessed the performance of the models by comparing the final grades obtained in this study with the grades students actually have.

IV. Results

A. Data Analysis of Student success

Data analysis of student success involves examining various factors and variables to gain insights into students' academic performance and achievement. It typically includes analyzing data related to students' demographics, prior knowledge, test scores, attendance, and other relevant information. By applying statistical techniques and visualization tools, we can identify patterns, trends, and correlations within the data, providing valuable information for educational institutions to make informed decisions and develop targeted interventions to support student success.

In order to perform predictions using the collected dataset, we employed the Python programming language, along with a selection of essential libraries. These libraries were seamlessly integrated within the Google Colaboratory, a no-cost cloud-based environment designed for executing Jupyter notebooks and securely storing them on Google Drive.

The data we extracted and utilized in our project consisted of student id, lessons, tests, student gender, the count of successful and unsuccessful tests, is the student repeating the course or not, midterm exams, the grade at the end of the course, and has the student successfully completed the course or not. For clarity, we associated each lesson with its corresponding test.

The gender-based disparity in test performance is a notable observation from our dataset, as depicted in Figure 2. Beyond just the sheer count of tests passed and failed, it's essential to delve deeper into the underlying factors contributing to this distinction. Understanding why females achieved higher success rates in passed tests and whether there are specific challenges or opportunities associated with each gender's academic journey could inform strategies for creating a more inclusive and equitable learning environment. This valuable insight can guide educators and institutions in providing tailored support and interventions, ultimately fostering academic success for all students.

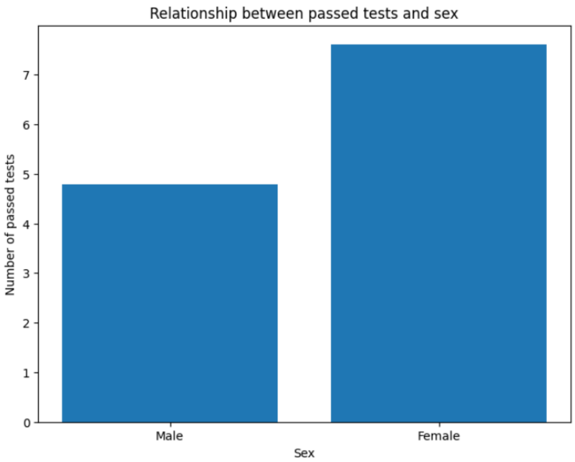

Fig. 2. *The ratio of passed tests by gender*

The graph displayed in Figure 3 illustrates a comparison between the average number of failed tests for men and women. In this visual representation, two bars are presented, with the x-axis denoting gender categories labeled as ’Male’ and ’Female,’ and the y-axis representing the average number of failed tests. The data depicted in Figure 3 clearly indicates that men tend to have a higher average number of failed tests in comparison to women. This observation sheds light on the gender-based disparities in test performance within the dataset.

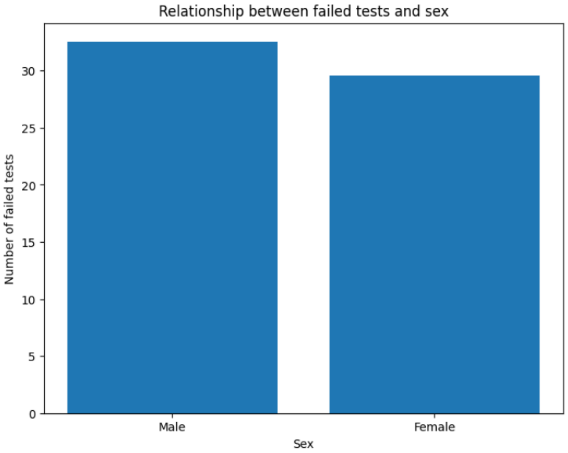

Fig. 3. The average of failed tests by gender

The next step involved calculating the average pass rates for the lessons and the average test scores. In the average scores for each test the best results were achieved in the tests on fundamental programming concepts, Python language elements, and loop structures with condition checking.

The mentioned tests were conducted in the first part of the semester and yielded results exceeding 35%, which are, concurrently, the best results among all tests. Slightly lower, but still very good results are also demonstrated by the tests on numerical data types (34%) and logical data types (33%). The tests for later lessons performed significantly worse, with the worst results observed in tests on recursion, debugging, error detection, and correction (below 1%). Slightly better but still poor results were obtained in sorting tests (Insertion, selection, quick sort) and file tasks (File Tasks - Numeric Issues, File Tasks - String Problems), where both group of tests yielded between 3% and 5% scores.

The subsequent step entailed the computation process of average completion rates of the lessons available to students on Moodle. The lessons on textual and logical data types demonstrate the best results. The lesson on textual data types yielded results exceeding 31%, which is, the best result. Next, the lesson on logical data types yielded a somewhat lower outcome of 29%. Slightly lower completion rates are observed for lessons on fundamental programming elements (27%) and Python language elements (25%). The lowest completion rates are seen in lessons on basic algorithms - string problems, digit problems, number divisor problems, Insertion sort, file-related problems - digit problems, and as evident from the test results, lessons on recursion and debugging. All of these lessons provided less then 1% results.

In Figure 4, we can observe the correlation between gender and prior knowledge for individual students. The prior knowledge variable was derived by averaging the scores of the initial test and 6 flash tests. This calculation resulted in an average value for each student. The graph clearly demonstrates that women attained a higher average score in prior knowledge compared to men, with their results hovering around 50%, whereas men scored below 40%.

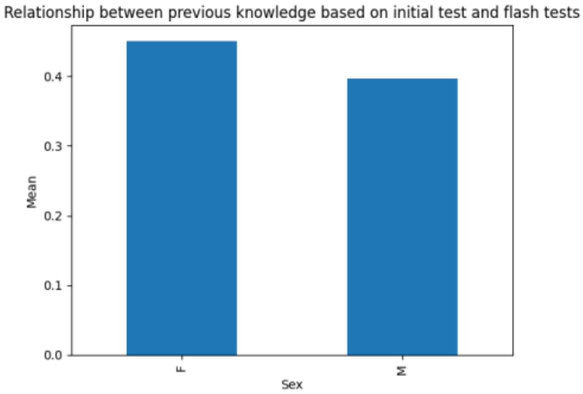

Fig. 4. The prior knowledge based on gender

The Figure 5 displays the distribution of students based on their final grades. As evident from the graph, the results obtained depict a downward curve, with the majority of students failing to meet the requirements of the course, and only a small proportion of students achieving an excellent grade.

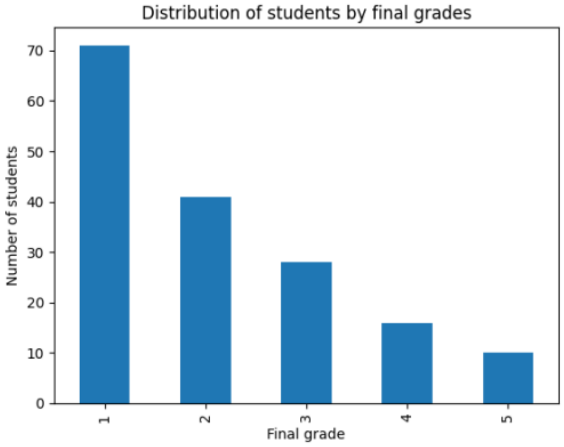

Fig. 5. The distribution of students based on their final grades

B. Correlation Analysis

Additionally, by employing a data plot for visualization, we graphically displayed coefficients of correlation, which quantify the relationships among variables. Consequently, we determined the variables that exhibit statistical relationships, irrespective of whether they have a causal connection or not. Due to the large number of variables in our dataset, we have chosen to look the correlation matrix for only the top 10 variables that exhibit the highest correlation.

The variables that showed the highest correlation were: Lesson Fundamentals of Programming - Lesson Basic Algorithmic Structures (correlation 0.71), Tests: Basic Algorithms (Digit Problems) - Lesson: Basic Algorithms (Digit Problems) (correlation 1). On the other hand, the variables Lesson: Basic Algorithms (Digit Problems) - Test: Built-in Function Test exhibited the lowest correlation of -0.09. As expected, there is a strong correlation between tests and their corresponding lessons, while there is a weaker correlation between lessons and tests from the beginning of the semester compared to those from the end of the semester. The correlation decreases as the semester progresses, and that is why lessons from the beginning of the semester have a negative correlation with lessons from the end of the semester, as more advanced concepts are being taught.

C. Predictive models

We have decided to utilize two models for prediction. The first model aims to predict is a student has successfully completed or not the course, while the second model aims to predict the final grade achieved by the student. To ensure accurate predictions for the first model which predicts whether a student has successfully completed the course, we first examined the correlations among all the data and removed any variables that showed no correlation. Following that, we divided the dataset into separate training and testing subsets and then proceeded to train multiple algorithms to identify the most efficient prediction model.

After that we applied data scaling, Principal Component Analysis (PCA), and generated new datasets that are scaled and reduced to a smaller number of attributes. Data scaling helps to align the range of attribute values, while PCA is a statistical technique used to simplify and explore high-dimensional data. This method is a dimensionality reduction technique that identifies the most significant features or variables within a dataset and expresses them as a new set of uncorrelated variables referred to as principal components. These principal components are linear combinations of the original variables and are arranged in order of the variance they account for in the data. By retaining a subset of the principal components that capture the majority of the variance, PCA allows for data visualization, pattern recognition, and exploratory data analysis. It is commonly used in various fields, including data science, image processing, genetics, and social sciences, to gain insights and interpret complex data structures. This approach offers several benefits, including speeding up machine learning algorithms, removing noise and unnecessary information from the data, and enabling data visualization in a lowerdimensional space.

Since the initial data set was insufficient for the second model to learn effectively, a new data set was created specifically for the second model. This new data set included variables such as "midterm exam1," "midterm exam2," "Final grade," "Repeater" (indicating whether the student repeated the year or not), "Number of passed tests," "Number of failed tests," and "Sex." To ensure data integrity, any missing values in the "midterm exam1" and "midterm exam2" columns were identified and replaced with zeros. The target variable for prediction was the "Final grade," which was subsequently removed from the data frame. Following that, the dataset was devided into separate training and testing subsets, and machine learning algorithms similar to those used in the first model were applied.

In the Table 1, we have presented Accuracy values of all evaluated models in prediction of student success in passing or failing a course. The Gradient Boosting Classifier model achieved an accuracy of 96.296%. This indicates that this model is highly effective in accurately predicting the outcome of a student passing or failing a course. The Random Forest model showed relatively high accuracy of 92.593%. The Neural networks model, which achieved an accuracy of 88.889% and Decision Tree model with 85.185% of an accuracy maybe are not perfect, but they still performed well in distinguishing between passing and failing. The Logistic Regression and Support Vector Machine models had moderate accuracies of around 66.667% respectively, suggesting that they have some predictive capability but are less accurate compared to the aforementioned models. Stochastic Gradient Descent and GaussianNB demonstrated lower accuracies ranging from 51.852% to 59.259%, indicating varying levels of effectiveness in predicting the pass/fail outcome.

TABLE I
THE ACCURACY OF PREDICTING STUDENT SUCCESS OR FAILURE IN A COURSE

Model	accuracy
GaussianNB	59.259
BernoulliNB	55.556
LogisticRegression	66.667
RandomForestClassifier	92.593
SupportVectorMachine	66.667
DecisionTreeClassifier	85.185
KNeighborsClassifier	77.778
GradientBoostingClassifier	96.296
Stochastic Gradient Descent	51.852
Neural Networks	88.889

In the first prediction, we utilized an extensive set of features for model training. To reduce the data dimensionality, we applied Principal Component Analysis (PCA). Since the feature data is represented through dimensions, the display of the results did not include the tree of the most significant variables.

Due to the aforementioned dimensionality and the fact that three models (Random Forest, Gradient Boosting, Decision Tree) performed the best in prediction, the most significant variables for each individual algorithm were not displayed.

In the Table 2, we have presented Accuracy values of all evaluated models in predicting final grades of students. The Random Forest Classifier, the Decision Tree Classifier, and the Gradient Boosting Classifier models performed with an excellent accuracy of 92.593%, indicating their effectiveness in predicting final grades. The Neural Networks model achieved a relatively high accuracy of 74.074%. This suggests that this model is reliable in predicting final grades. The K Neighbors Classifier and GaussianNB models showed decent accuracies of 62.963%, implying their ability to provide reasonably accurate predictions. The Logistic Regression and BernoulliNB models had accuracies of approximately 59.259%, indicating moderate performance in predicting final grades. The Stochastic Gradient Descent, Support Vector Machine models exhibited lower accuracies ranging from 37.037% to 55.556%, suggesting limited effectiveness in predicting final grades.

TABLE II
THE ACCURACY OF THE MODEL FOR PREDICTING FINAL GRADES OF STUDENTS

Model	accuracy
GaussianNB	62.963
BernoulliNB	59.259
LogisticRegression	59.259
RandomForestClassifier	92.593
SupportVectorMachine	55.556
DecisionTreeClassifier	92.593
KNeighborsClassifier	62.963
GradientBoostingClassifier	92.593
Stochastic Gradient Descent	37.037
Neural Networks	74.074

Figure 6 presents the distribution of students based on their grades, providing a visual representation of how students are distributed across different grade categories. There is another visualization showcasing the distribution of students by grades, but this time based on predictions made by the Random Forest algorithm. The Random Forest algorithm has been identified as one of the best-performing model during the validation of the results. This visualization demonstrates how well the Random Forest algorithm categorizes students into different grade categories.

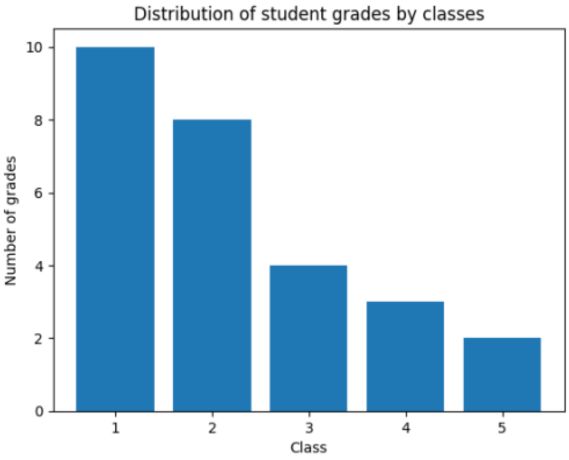

Fig. 6. Random forest - distribution

As Random Forest algorithm has been recognized as one of the top-performing model, in the visualization below Figure 7, we can observe the outstanding performance. The results of the confusion matrix over the test data are presented. The confusion matrix serves as a comprehensive assessment of the model’s predictive capabilities by assessing the accuracy of its predictions for each class.

In the model for the prediction of students final grades, we employed a different dataset that allowed us to showcase the most important variables as a tree. We extracted the first four most significant variables in the model, which are: midterm exam1, midterm exam2, number of passed tests, and number of failed tests. We visualized the variables as a tree for one of the best performed model, in our case it was the Random Forest model.

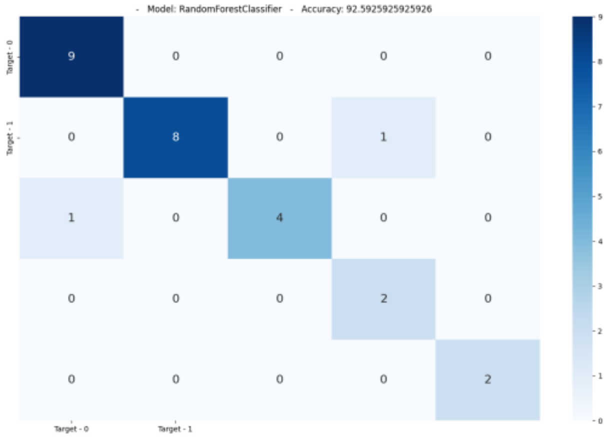

Fig. 7. Random forest - confusion matrix (test)

In summary, the Random Forest Classifier, Decision Tree Classifier, Gradient Boosting Classifier, and Neural Networks models consistently performed well across both prediction tasks. These models may be the most reliable options for predicting whether a student achieved success or faced failure in a course and for estimating their final grades.

D. Mathematical model

In this section, we explore the mathematical foundations of three key predictive models: Decision Tree, Random Forest, and Gradient Boosting, used in predicting student performance within the Moodle Learning Management System (LMS).

The Decision Tree model constructs a tree structure based on nodes and branches. The DT algorithm is part of the supervised learning algorithm family, and its main objective is to construct a training model that can be used to predict the class or value of target variables through learning decision rules inferred from the training data. Each node represents a test on a specific feature, and branches segregate data based on the outcomes of these tests. Key matematichal concepts for the model are Entropy and Information Gain.

Entropy measures the impurity or disorder in a set of data.

$$E(S) = \sum_{I=1}^{C} - p_i \log_2 p_i$$

where pi is ratio of the sampe number of the subset and the i-th attribute value [14].

Information Gain quantifies the effectiveness of a particular feature in reducing uncertainty. The data gain Gain(S,A) is defined as shown:

$$Gain(S, A) = \sum v \epsilon V(A)|S||Sv|Entropy(Sv)$$

where the range of attribute A is V(A) and Sv is subset of the set S qual to the attribute value of the attribute v.

Random Forest, an ensemble learning technique, enhances predictive accuracy by aggregating the results of multiple decision trees. The key mathematical concepts underpinning this method include bagging (Bootstrap Aggregating) and averaging. Bagging involves the random sampling of instances with replacement, creating diverse datasets for each decision tree within the ensemble. This process introduces variability and mitigates overfitting by ensuring that each tree is trained on a distinct subset of the data. Averaging is employed to combine the predictions from individual trees, thereby reducing the impact of potential overfitting in any single tree and improving the model's generalization capabilities. The algorithmic steps of Random Forest further elucidate its functioning. The process begins with bootstrapping, where the training data is randomly sampled with replacement to create multiple distinct training sets for each decision tree. Subsequently, the algorithm proceeds to build decision trees using different subsets of the data. These trees are constructed through a recursive process, selecting optimal features at each node based on criteria such as Gini impurity or information gain. The construction of diverse trees contributes to the overall robustness of the ensemble. Finally, predictions $\hat{Y}$ are obtained by averaging the outputs of individual trees by employing a majority vote for classification task [15].

$$\hat{Y}_{RF} = \frac{1}{N} \sum_{i=1}^{N} \hat{Y}_i$$

Gradient Boosting, an ensemble technique, iteratively builds weak learners, often Decision Trees, to correct errors from previous iterations. Key mathematical concepts include Gradient Descent, optimizing the model by adjusting parameters toward the steepest descent of the loss function, and the use of Weak Learners, simple models like shallow trees, combined for a strong predictive model. The algorithm involves the iterative construction of weak learners to rectify errors, applying Gradient Descent to adjust data point weights, and combining models by summing weak learners for a robust predictive model. The model minimizes the loss function L with respect to the predictions f(x):

$$\text{minimize} \ \sum_{i=1}^{N} L(y_i, \ f(x_i))$$

where L is a suitable loss function, and y_i is the true label [16].

V. Discussion

In recent years, the research on e-learning has experienced significant expansion, with the utilization of tools and methods to understand student behavior in learning management systems (LMS) being integral aspects. Online learning systems have played a vital role in enhancing critical thinking abilities and introducing innovative approaches to course mastery among higher education students. The availability of reusable and adaptable learning materials has been expected to contribute to students' academic achievements. Therefore, it has been recommended the suggestion is to prioritize tasks that are intuitive and visually appealing. To better monitor and predict student performance, the use of specialized JavaScript-based applications has been suggested as they significantly reduce the time required for processing substantial data volumes during midterm and final exam. Traditional teaching approaches have lost their relevance. Consequently, there has been a need to develop progress oriented training courses that can foster students motivation not only for academic achievement but also for self-improvement. Educators have been encouraged to allocate tangible assignments using virtual educational platforms to foster the development of student competencies.

The widespread adoption of e-learning has grown due to its capacity to accumulate extensive information, analyze student behavior, and assist teachers in identifying potential errors during the learning process. Research carried out on the observational findings of applying various data mining algorithms utilized to explore the educational and predicting academic success has shown that features that can influence student achievement can be categorized into a smaller number of categories. These categories include factors such as demographic data, indicators of previous performance, course and instructor data, and general student data. These factors are unrelated of individual student personalities. Additionally, personal motivation factors are also important. In the context of an online learning environment, student enthusiasm has significantly influenced the excellence of the learning process. Data mining methodologies have provided valuable insights for appraising student incentive and optimizing the educational endeavor. One driver of student motivation has been contribution in collaborative projects, which promote teamwork and enhance interest in completing complex tasks. This encourages students to engage and communicate within the LMS framework.

The five main questions that were the focus of this research are listed in the introduction of this article. The question arises, what are the answers to these questions?

The first thing we were interested in is how effective learning analytics are in predicting student success in Moodle LMS. This study has shown that learning analytics can be effective in predicting student success in Moodle LMS. By analyzing various collected student data within the learning management system (such as test scores and lesson scores), learning analytics can offer a deeper understanding of student actions and trends. These insights can be used to identify students at risk and those who are not at risk when it comes to final grades in order to improve student outcomes. Learning analytics algorithms can uncover patterns and correlations between different data, enabling teachers to make informed decisions and interventions based on predictive models. However, it is important to note that the effectiveness of learning analytics in predicting student success depends on several factors. The quality and accuracy of the collected data, the sophistication of the models and algorithms used in analytics, and contextual factors within the Moodle LMS environment can impact predictive accuracy. Additionally, ethical and privacy considerations should be taken into account when using learning analytics. Overall, learning analytics in Moodle LMS has the potential to be extremely proficient in predicting student success, but its success relies on data quality, the analytics models used, and the context in which it is implemented.

The second question, which we are interested in answering, relates to the accuracy of the predictive models built using learning analytics in this research. In Table 1 and Table 2, there are presented ten models that were used in this research paper, and their corresponding accuracy. As it is shown in the mentioned tables, three models stand out with the highest accuracy for both predictions, and those models are: Random Forest Classifier, Decision Tree Classifier, and Gradient Boosting Classifier. An accuracy from 90% to 96% is considered satisfactory for the context and requirements of this study. However, it is important to emphasize that for this study we used a small dataset, which can indeed have an impact on the accuracy of the models. When working with limited data, the models may struggle to capture the underlying patterns and generalize well to new instances. This can result in less reliable predictions and potentially lower accuracy scores. To improve the accuracy and reliability of the models, it is advisable to gather a larger dataset that encompasses a wider range of examples, ensuring a more comprehensive representation of the target problem. Additionally, employing techniques such as cross-validation and regularization can help mitigate the impact of limited data and improve the models’ generalization capabilities.

Next, we were interested in how do different factors, like gender or previous academic performance, influence the effectiveness of predictive models in Moodle LMS. After conducting the research and analyzing the data on student success, it was found that women outperform men in relation to their prior knowledge, as measured by the results of the short unannounced tests. This difference in performance between genders was visually presented in the previous section, "Data Analysis of Student Success." These findings indicate a possible connection between gender and academic achievement, emphasizing the importance of further research and interventions aimed at understanding and supporting the specific needs and challenges that may arise in the context of gender and education.

The final question that was the focus of this research pertains to whether learning analytics in Moodle LMS can be used for personalization and adaptation of instructional approaches to meet the needs of individual learners. This is a question that may never be definitively answered, but it can be assumed that it will never be able to cater to all students and their needs because currently, there are no significant indications that a solution will be developed in the near future to accommodate everyone and all their needs.

Future studies on student success can encompass a extensive variety of factors that influence student achievement, involve a larger number of students, include a greater variety of courses taken by students, incorporate diverse populations, and so on.

VI. Conclusion

This study focused on preprocessing data from the "Programming 1" course at the University of Split to predict student outcomes. Challenges included ensuring assessment authenticity and handling missing data, impacting model performance. Objectives were to predict final grades and exam outcomes, aiming to increase online exam supervision and develop oral exam prediction models. To improve understanding of individual progress, activities like group assignments were proposed. Utilizing Moodle data for predictions could benefit teachers and students by identifying impactful attributes for success. However, addressing factors like teaching experience and technological access is crucial to improve engagement with e-learning tools. Our research built on Y. Zhang et al.'s work, emphasizing data preprocessing for predictions. Ten models were used, with Gradient Boosting Classifier showing the highest accuracy (96.296%) in predicting course pass/fail outcomes. Random Forest Classifier, Decision Tree Classifier, and Gradient Boosting Classifier consistently outperformed others, excelling in various evaluation metrics. Their strength lies in handling complex relationships between features and capturing key interactions, providing robust predictions despite missing data. However, the choice of the best model may depend on specific requirements, constraints, interpretability requirements, training time, and available data. The study confirmed that student conduct on online educational platforms has an impact on student performance. However, despite the increasing usage of these platforms, they do not provide enough information on their own to accurately predict students’ future outcomes. The findings demonstrated that without additional attributes that are currently unmeasurable by such systems, high-level predictions cannot always be achieved.

References

[1] Sunil, & M. Doja (2017). *Data mining techniques to discover students' visiting patterns in e-learning resources*. International Journal of Computer Science and Mobile Computing (IJCSMC), 6, 363–368.

[2] R.S. Baker, T. Martin, L.M. Rossi (2016*). Educational data mining and learning analytics.* The Wiley handbook of cognition and assessment: Frameworks, methodologies, and applications, pp. 379–396.

[3] R.S. Baker, K. Yacef, et al. (2009). *The state of educational data mining in 2009: A review and future visions*. Journal of Educational Data Mining, 1(1), 3–17.

[4] R. Conijn, C. Snijders, A. Kleingeld, U. Matzat (2016). *Predicting student performance from LMS data: A comparison of 17 blended courses using Moodle LMS*. IEEE Transactions on Learning Technologies, 10(1), 17–29.

[5] E.B. Costa, B. Fonseca, M.A. Santana, F.F. de Araújo, J. Rego (2017). *Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses.* Computers in Human Behavior, 73, 247–256.

[6] D. Gašević, S. Dawson, T. Rogers, D. Gasevic (2016). *Learning analytics should not promote one size fits all: The effects of instructional*

conditions in predicting academic success. The Internet and Higher Education, 28, 68–84.

[7] L.P. Macfadyen, S. Dawson (2010). *Mining LMS data to develop an "early warning system" for educators: A proof of concept*. Computers & Education, 54(2), 588–599.

[8] M. Riestra-González, M. del P. Paule-Ruíz, F. Ortin (2021). *Massive LMS log data analysis for the early prediction of course-agnostic student performance.* Computers & Education, 163, 104-108.

[9] C. Romero, P.G. Espejo, A. Zafra, J.R. Romero, S. Ventura (2013). *Web usage mining for predicting final marks of students that use Moodle courses*. Computer Applications in Engineering Education, 21(1), 135–146.

[10] G. Siemens (2013). *Learning analytics: The emergence of a discipline*. American Behavioral Scientist, 57(10), 1380–1400.

[11] G. Siemens, R.S. J. d. Baker (2012). *Learning analytics and educational data mining: Towards communication and collaboration*. In Proceedings of the 2nd International Conference on Learning Analytics and Knowledge (pp. 252–254).

[12] N. Tomasevic, N. Gvozdenovic, S. Vranes (2020). *An overview and comparison of supervised data mining techniques for student exam performance prediction*. Computers & Education, 143, 103-676.

[13] Y. Zhang, A. Ghandour, V. Shestak (2020). *Using learning analytics to predict students' performance in Moodle LMS*.

[14] B.T. Jijo, A.M. Abdulazeez (2021). *Classification Based on Decision Tree Algorithm for Machine Learning*. Journal of Applied Science and Technology Trends, 02(01), 20–28.

[15] G. Biau, E. Scornet (2016). *A Random Forest Guided Tour*. Test, 25, 197–227.

[16] C. Bentéjac, A. Csörgő, G. Martínez-Muñoz (2021). *A Comparative Analysis of Gradient Boosting Algorithms*. Artificial Intelligence Review, 54, 1937–1967.

Mia Mužinić is a student at the Faculty of Science at the University of Split, currently in her final year pursuing a Master's degree in Data Science and Engineering, having completed three years of undergraduate studies in Computer Science. Alongside her academic pursuits, she engages as an educator, teaching programming courses at Algorithmics, focusing on imparting coding skills to children. While pursuing her Master's degree in Data Science and Engineering, Mia has actively contributed to academic research. She dedicated her efforts to a scientific paper and a project focusing on Deep Learning approaches for Video Classification.

Antonela Sikavica is a student at the Faculty of Science at the University of Split, currently in her final year pursuing a Master's degree in Computer Science with a teaching specialization. She has previously completed her undergraduate studies in Computer Science. Alongside her academic pursuits, she actively works as a computer science teacher in both an elementary school and at Algorithmics, teaching programming to children. Her commitment to education is evident through her dual roles, aligning her practical teaching experiences with her ongoing academic journey.

Petra Zelić recently completed her Master's program in Computer Science with a teaching specialization at the Faculty of Science, University of Split. Prior to this achievement, she concluded her undergraduate studies in Computer Science at the same institution. She currently holds the position of System Analyst at Privredna banka Zagreb d.d. Prior to this role, she gained experience during the summer of 2023 as a Data Science and ML Developer Intern at Ericsson Nikola Tesla d.d. where she participated at the SoftCom conference 2023 with her team members. Petra also engaged as a Demonstrator for the Programming 1 course at the University of Split, Faculty of Science, from October 2022 to February 2023. Additionally, Petra held the role of Demonstrator – Junior Dev React.JS at the digital agency Digitalna Dalmacija from March to May 2023. Petra also participated in preparing students for the competition in the Algorithms category as an External Mentor at the Faculty of Science. Her extensive education, work experience, and diverse skill set position her as an expert ready to make further contributions to the field of Computer Science.

Ani Grubišić is an Associate Professor at the Faculty of Science at the University of Split, a faculty member since 2002, teaching Introductory Programming in Python, E-learning Systems and Learning Analytics courses. She is a principal investigator of her second Office of Naval Research grant awarded to a research project Enhancing Adaptive Courseware Based on Natural Language Processing. She has memberships in the IEEE Technical Committee on Learning Technology (IEEE TCLT), the IEEE Adaptive Instructional Systems Technical Advisory Group (IEEE AIS TAG), and the IEEE Adaptive Instructional Systems Standards Working Group (IEEE AIS WG). She is also the Editorial Board member of the International Journal of Educational Technology in Higher Education (ETHE), Associate Editor of the Journal of Communications Software and Systems, Program Board Member for the International Conference on Adaptive Instructional Systems (AIS/HCI), Chair for the Symposium on Advanced Educational Technologies (part of International Conference SoftCOM), and a reviewer for the International Journal of Educational Technology in Higher Education, the Education and Information Technologies, the IEEE Transactions on Learning Technologies, the Artificial Intelligence Review, and the Computers & Education.

Ines Šarić-Grgić is a Research Assistant at theFaculty of Science, University of Split, Croatia,and a PhD student of Computer Science at theFaculty of Electrical Engineering, MechanicalEngineering, and Naval Architecture, Universityof Split, Croatia, where she earned her MSc in2008. In 2012, she received an MScspecialization in Business Economics at theFaculty of Economics, University of Split,Croatia. Since 2015, she has worked on two projects funded by the Office ofNaval Research, USA. Her research interest includes artificial intelligence ineducation and intelligent tutoring systems.