IRSTI 28.23.29 https://doi.org/10.26577/jpcsit2023v1i4a6

G.H. Aimal Rasa[1] , Z.M. Abdiakhmetova[2*]

[1]Kabul Education University, Kabul, Afghanistan
[2]Al-Farabi Kazakh National University, Almaty, Kazakhstan
*e-mail: zukhra.abdiakhmetova@gmail.com

STUDY OF SIGNS OF IMPACT ON THE QUALITY OF EDUCATION BY ML

Abstract. Use of machine learning (ML) algorithms to analyze and identify signs that affect the quality of education opens up new opportunities for individualization of education, optimization of educational processes and improvement of educational results of students.

The personalization of education stands as a paramount trend in contemporary learning. Each student possesses distinct requirements, passions, and talents. By scrutinizing the factors impacting educational excellence, we can pinpoint individual elements that wield substantial influence over each student's success. Consequently, this enables the crafting of customized educational programs and techniques finely tuned to the unique needs and aptitudes of every student.

The objective of this endeavor is to construct a system employing machine learning algorithms that can discern the factors influencing the assessment of students' educational quality. These facets render the research pertinent and noteworthy within the landscape of modern education. It offers a gateway to a deeper comprehension of the learning processes, streamlining educational procedures, and ultimately yielding improved outcomes in student education.

Key words: Machine Learning Algorithm, Support Vector Method, Random Forest, Dataset, Linear Regression.

1. Introduction

Currently, the process of digitalization of society is developing very quickly. This, in turn, simplifies many tasks encountered in everyday life. Of course, this phenomenon finds its place in education area and affects the development of this field. Lydia Sandra, Ford Lambangaol, and Tokuro Matsuo [1] reviewed the factors that may affect student performance and technologies that help predict student performance.

The study delved into several inquiries, one of which aimed to identify key attributes for forecasting student achievement. The findings revealed that both internal assessment and summative assessment emerged as the most consistent indicators for predicting academic performance. Additionally, the study highlighted other significant factors such as personal and self-assessment, prior academic records, extracurricular involvement, and social characteristics.

One of the primary objectives of this research was to explore the application of machine learning algorithms in forecasting student performance. The study was conducted in May 2021, and the data was sourced from the IEEE Access and Science Direct databases, adhering to standard database search protocols, including exclusion and inclusion criteria and search result analysis. The study employed a range of algorithms, including logistic regression, Bayesian methods, K-nearest neighbor (KNN), regression trees, random forest, decision trees, long-term and short-term memory (LSTM), support vector machines (SVM), multi-layer perceptron neural networks (MLP), artificial neural networks (ANN), and reference vector methods, with all of them demonstrating promising results [2].

Another research study on this topic [3] highlights the challenge of predicting student performance due to the vast volume of data in educational databases. The primary objective of this research is to review the use of artificial intelligence systems for predicting academic learning.

The research paper presents an automated system for evaluating students' progress and analyzing their achievements. Here, the author uses a tree algorithm to accurately predict student performance. The data clustering method was used to analyze a large set of student databases. This method speeds up the search process and provides accurate classification results. A new model of teaching is proposed using information about a student obtained during college registration. The final data sets are entered so that Machine Learning algorithms can use them and predict student performance. During the research, 13 algorithms were selected from Machine learning algorithms in 5 categories: Bayesian, SVM, MLP,

IBK, linear regression and tree-type algorithms. As a result of the study, the use of binomial logistic regression gave 97.06% accuracy, decision tree 88.24%, entropy 91.18%, K-nearest neighbor showed 93.72% accuracy. Among the machine learning algorithms, binomial logistic regression performed best.

The research paper in question [4] emphasizes that a nation's economic prosperity hinges on the accessibility of higher education, a concern at the forefront of any government's agenda. Hussein Altabrawi's insights further underscore that the surge in student loan debt in America is partly attributable to delayed graduation rates. To construct the machine learning model, a fully connected artificial neural network, Bayesian algorithm, decision tree, and logistic regression were employed.

The dataset used to develop these models was compiled from students at the College of Humanities during the 2015 and 2016 academic years, sourced from a combination of student surveys and test books. This dataset encompasses information pertaining to 161 students. The research activities encompassed the creation of a student dataset, data collection, data preprocessing, the construction and evaluation of four machine learning models, identification of the best-performing model, and a thorough analysis of the results.

The dataset is comprehensive in nature, encompassing factors ranging from the students' age and gender to the extent of family involvement in their education. The effectiveness of these four models in relation to the dataset was assessed using the ROC index, with the artificial neural network achieving a value of 0.807, Bayesian at 0.697, decision tree at 0.762, and logistic regression at 0.767. Notably, the artificial neural network emerged as the most effective model [5].

2. Literature review

The next work [6] was conducted in order to develop and compare different Machine Learning algorithms to predict students' academic achievement. The work was published in the journal Computers & Education and is widely recognized in the field of education and information technology.

The work used the data of students with the following characteristics:

- Data on student demographics such as age, gender, mother tongue and nationality;
- Information about the activity of students on the learning platform, for example, the number of visits, the duration of sessions and the time spent on tasks;
- Results of tests and assignments, including scores, quizzes, work grades, and number of correctly solved assignments.

Another work in this direction [7] describes the use of various machine learning methods to predict student achievement. The paper describes the various indicators that can be used to predict student performance, as well as the advantages and disadvantages of using machine learning in this field.

In their article [8], authors Muneer Ahmad Al-Radaideh, Ghassan Issa, and Mohammed Al-Zyoud employed their dataset to forecast the academic performance of students enrolled at a Jordanian university. To achieve this objective, they applied various machine learning algorithms and determined the most efficient one for this specific task.

The outcomes of their research demonstrated that Machine Learning techniques can indeed be highly effective in predicting students' academic success. The researchers employed several Machine Learning algorithms, including decision trees, Bayesian classifiers, and the support vector method (SVM), and conducted a comparative analysis of their performance. The results revealed that the SVM algorithm outperformed the others, achieving a prediction accuracy rate of over 90%. Moreover, the researchers identified key factors such as prior exam GPAs, student age, the number of subjects studied, and the presence of family support as the most influential in determining academic achievement. Consequently, this study provides further evidence of the feasibility of employing machine learning methods to predict students' academic success and pinpoint significant factors that impact their educational accomplishments.

Researchers in their work [9] used machine learning methods to predict students' final grades. To do this, they used data on the academic performance of 600 students at a Portuguese university. Information on socio-economic characteristics such as Student's study schedule, number of absences, number of study materials available to the student, age and gender of the student, as well as parents' education and access to the Internet at home were used as indicators of analysis. The researchers used several machine learning algorithms, including logistic regression, random forest, and gradient boosting, to predict students' final grades. They also compared the effectiveness of machine learning techniques with traditional methods such as linear regression and multiple regression.

The results showed that machine learning methods are more effective in predicting students'

final grades than traditional regression methods. The researchers also found that the most important predictors of student achievement were internet access at home, the student's study schedule, and absenteeism.

In a research paper [10], scholars conducted a comparative investigation of various machine learning algorithms for the purpose of forecasting student performance. The study employed a dataset comprising the academic records of 649 engineering students at an Indian university. The dataset included information regarding students' online learning activities, encompassing metrics such as the number of lectures viewed, tasks completed, responses to test questions, and the average time taken to answer those questions. Additionally, the study considered background characteristics such as students' age, gender, and geographic location as factors for analysis. The researchers assessed multiple machine learning algorithms, including logistic regression, multiple linear regression, decision trees, random forest, and gradient boosting, to predict student performance. They also compared the efficacy of these algorithms and identified the most crucial attributes for forecasting student success. The findings revealed that gradient boosting emerged as the most effective algorithm for predicting student performance, exhibiting robust performance across diverse data configurations.

In a separate study [11], researchers evaluated the effectiveness of different machine learning algorithms in predicting student performance using the «student performance dataset,» which contains data on student achievements in mathematics and Portuguese language courses. This dataset incorporates various aspects, including demographic information, academic background details, family-related data, and student behavior patterns.

Furthermore, the work presented in [12] constitutes a case study aimed at predicting the progress of engineering students through the application of Machine Learning algorithms. The authors gathered data from 235 engineering students attending a private university in India. This dataset encompassed a wide array of features, such as student demographic data, progress records, attendance records, and assignment grades. The researchers employed various Machine Learning algorithms, including decision trees, random forests, and support vector machines, to forecast student scores.

In the research conducted by M. Abdullahi and S. K. Abdullah in their work [13], they explored the use of different machine learning algorithms. Their study was based on data collected from the University's information system in Iran and aimed to predict student performance in higher education. Various indicators were utilized in their predictive models:

- Average number of hours spent studying per week.
- Average score in mathematics, physics, chemistry, English subjects at school.
- Student course at the university.
- The size of the student study group.
- Student status (for example, whether the student is an athlete or not).
- Status of the student in relation to military duty.
- The student's employment status (for example, whether the student is employed or not).
- The average score of the student in the previous semesters at the university.

The authors employed the following Machine Learning algorithms to forecast student performance: Random Forest, Gradient Boosting, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Decision Trees.

One of the primary strengths of this research lies in its utilization of real data extracted from the University's information system in Iran. This ensures that the findings have practical applicability and can be instrumental in enhancing the higher education process. An article by V. Nair, S. Bhatia, and D. Dey in 2021 [14] offers an overview of numerous studies utilizing machine learning techniques for predicting student performance. In their publication, the authors scrutinized 66 articles published from 2010 to 2020 and identified common indicators utilized across these studies to forecast student performance.

However, despite the significance of this research, it does have some limitations, notably a relatively modest feature set and a lack of comparisons between machine learning methods.

In another research paper [15], the authors conducted an extensive literature review to assess existing machine learning approaches for predicting student performance in educational institutions. They compared various machine learning techniques employed in research. The authors incorporated diverse indicators for forecasting student performance, encompassing demographic information, educational data, test scores, student attendance and engagement data, among others. The incorporation of such indicators enables more accurate predictions of student progress and the identification of factors influencing their academic accomplishments.

In [16], the authors proposed a machine learning model for predicting student performance using

student data. They explored various machine learning techniques, including decision trees, random forest, logistic regression, and the support vector method. The analysis utilized data from 183 students, encompassing various course-related characteristics such as age, gender, ethnicity, prior education's grade point average, school type, number of subjects, completion and duration of standardized math tests, and the number of lectures and exercises. The results demonstrated the effectiveness of machine learning models in predicting students' performance in higher education, particularly when employing algorithms like random forest and the support vector method.

Overall, this paper constitutes a valuable contribution to the field of utilizing machine learning models to predict student performance in higher education. However, it is essential to acknowledge some limitations and drawbacks of this work.

In an article by S. N. Tiwari and A. K. Misra [17], the authors explored the potential of machine learning techniques in predicting student performance in online courses (MOOCs). They considered four distinct machine learning algorithms: decision tree, random forest, nearest neighbor method, and logistic regression. Indicators derived from student activity on the MOOC platform, such as time spent in the course, assignments completed, forum participation, etc., were employed. Additional indicators, such as word count in task responses and video lecture views, were also utilized. The research utilized data from the course «Introduction to Computer Science and Programming Using Python» on the edX platform, involving a total of 1972 students. The results indicated that the random forest method achieved the highest accuracy with an F1 score of 0.85, while logistic regression yielded the lowest precision with an F1-measure of 0.72. The authors concluded that Machine Learning techniques can serve as an effective tool for predicting student performance in MOOCs and recommended the random forest approach for this task.

In the subsequent article [18], B. Kabir and H. Ali conducted a systematic review of literature concerning the utilization of machine learning algorithms for predicting student performance in higher education. Their review encompassed 65 articles published between 2015 and 2020. The authors utilized the following indicators to forecast student performance:

Demographic data (gender, age, race, etc.).

Academic data (grade point average, exam grades, credits, etc.).

Behavioral data (frequency of visits, time spent in the course, number of materials visited, etc.).

Social data (activity in social networks, communication with peers and teachers, etc.).

The authors compared various Machine Learning algorithms used for predicting student performance, including logistic regression, decision trees, support vector methods, random forests, and neural networks. They also assessed prediction quality using metrics such as accuracy, completeness, F-measure, and ROC curve.

3. Material and Methods

Collecting Data

In total, the data set contained 21 characters. Labels in the data set:

- Gender (binary: 1-male, 0-female)
- Age characteristics (number: from 15 to 17 years old)
- Is your school near your home? (binary: 1-yes, 0-no)
- Is your family complete? (binary: 1-yes, 0-no)
- Your mother's education?
- Your father's education? (numerical value: 1-medium, 2-unfinished higher, 3-higher, 4-higher post-graduate professional)
- Your mother's job? (binary: 1-employed, 0-unemployed)
- Do your parents do work that requires physical effort? (binary: 1-yes, 0-no)
- Your father's job? (binary: 1-employed, 0-unemployed)
- Is there internet connection at home? (binary: 1-yes, 0-no)
- What do you do in your spare time? (binary: 1-useful, 0-useless)
- Do you meet your friends often? (binary: 1-yes, 0-no)
- Do you eat well? (binary: 1-yes, 0-no)
- What is your phone's operating system? (binary: 1-Ios, 0-Android)
- Do you have a close relationship with students? (binary: 1-yes, 0-no)
- Are you satisfied with your social situation? (binary: 1-yes, 0-no)
- Can the teacher interest you in his subject? (binary: 1-yes, 0-no)
- Does the teacher evaluate fairly (competently)? (binary: 1-yes, 0-no)
- G1- previous term grades (number: 3 to 5 years)
- G2-prior term grades (numerical: 3 to 5 years)
- G3-estimated value (number: 3 to 5 years)

On the basis of the selected signs, a survey was conducted among the upper classes of the

specialized lyceum named after Al-Farabi. Data attributes include student grades, demographic, social, and school characteristics. These data were collected through school reports and questionnaires. Indicators include the age and gender of the students, the occupation of the parents, and a healthy daily diet. A total of 505 students participated in the survey. 234 of them are men, 271 are girls. In the dataset, the textual data were converted to binary and numeric data for use in calculations. Attributes with a binary value (gender, marital status, completeness of the family, parental work, frequent meetings with friends, internet access, healthy nutrition, teacher's interest in his subject and competent assessment) were changed to 0 and 1 according to the response values. Also, the education of the student's parents was numerically divided into 4 different values. They are as follows: 1-secondary, 2-incomplete higher, 3-higher, 4-higher post-graduate professional.

After transforming the data, we need to classify it. In the following program code, we divide the data into two parts: the labels and the target variable.

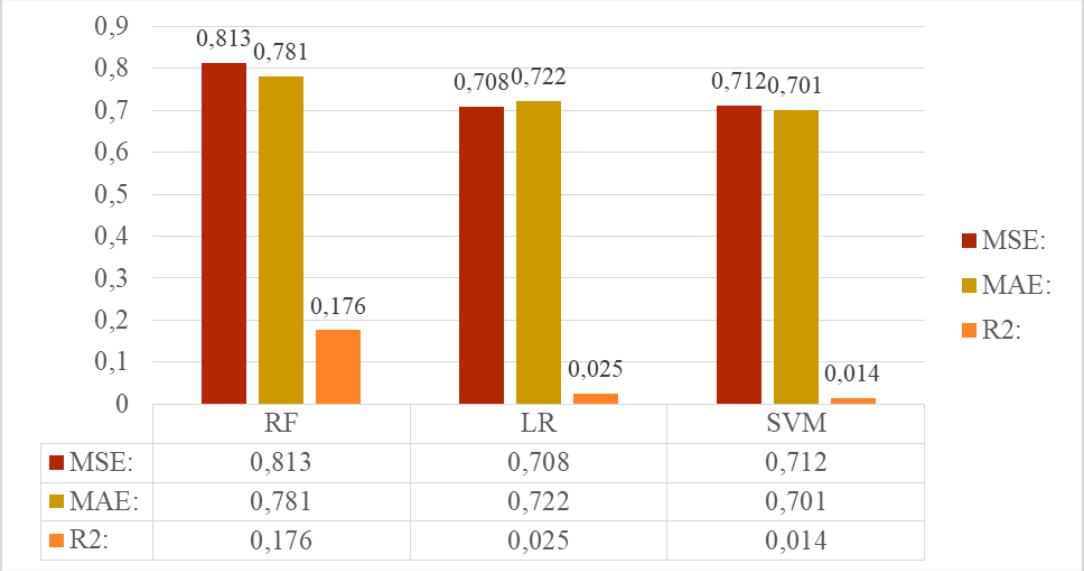
4. Processing of data collected during pedagogical practice using machine learning algorithms

After converting, classifying and grouping our data, we proceeded to our calculations using 3 selected algorithms. We used the Random Forest method for the first time. The next method used in the calculation is the linear regression method. The linear regression method is widely used in forecasting and classification. The advantage of the linear regression method over other methods is its simplicity and the results can be easily interpreted.

The last method used in forecasting is the method of reference vectors or the method of support vectors. This code is used to train a model using support vectors (SVM) on the training data (x_train and y_train).from sklearn.svm import SVC: This line of code imports the SVC (Support Vector Classifier) class from the SVM module into the scikit-learn (sklearn) library. SVC is used for classification problems using the reference vector method.

5. Analysis of the results obtained from the experiment

The accuracy of each algorithm was calculated using the above MSE (Mean Squared Error), MAE (mean Absolute Error) and R2 (R-squared) indicators. According to the results of these evaluation indicators, the mean square error of the linear regression method is 0.708, the mean absolute error is 0.722, R-squared is 0.025, the mean square error of the reference vector method is 0.712, the mean absolute error is 0.701, the R-squared is 0.014, the random forest method mean square error was 0.813, mean absolute error was 0.781, R-squared was 0.176. Accordingly, among the methods, the method of reference vectors showed the best results (Figure 1).



	RF	LR	SVM
MSE:	0,813	0,708	0,712
MAE:	0,781	0,722	0,701
R2:	0,176	0,025	0,014

Figure 1 – Comparing of methods

Also, among the symptoms relevant to the student, the prediction and the magnitude of the symptoms that affected the

Among the characteristics, characteristics such as the student's age, parents' education, free time and grades of the previous term were found to be more influential than others.

The main features were found to be the student's age, parents' education, free time and grades of the previous term (Figure 3).

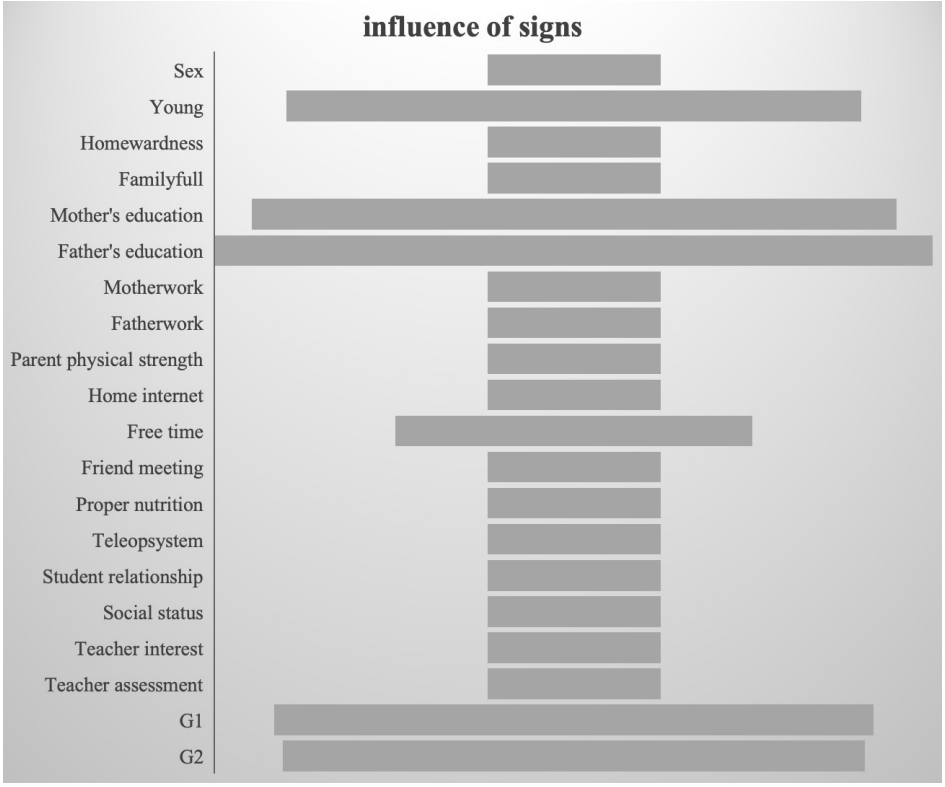


Figure 2 – Influence of Signs

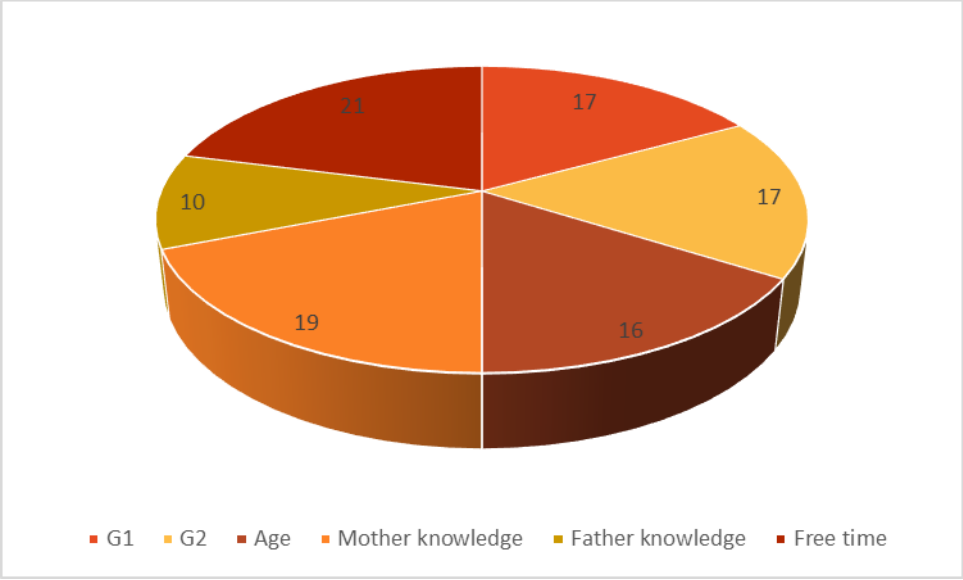


Figure 3 – Main Features of dataset

The selection of these signs has its own reasons and features.

6. Conclusion

Throughout the course of our investigation, we examined the factors influencing the quality of students' education. Our initial objective involved identifying the characteristics relevant to the utilization of machine learning algorithms within the realm of education. We conducted comprehensive literature reviews to explore existing studies that had employed machine learning algorithms in educational contexts. Subsequently, we engaged in research and analysis of these machine learning algorithms.

In the practical part of the research work, a survey was conducted on 18 signs in Al-Farabi specialized lyceum. The second task was to use symbols and machine learning algorithms in calculations. Calculations were made on the data set collected on the basis of the survey by the method of linear regression, random forest, and support vectors. Students' grades were estimated. We used MSE (Mean Squared Error), MAE (mean Absolute Error) and R2 (R-squared) indicators to assess the accuracy of the methods.

According to the evaluation, the mean square error of the linear regression method is 0.708, the mean absolute error is 0.722, R-squared is 0.025, the mean square error of the reference vector method is 0.712, the mean absolute error is 0.701, R-squared is 0.014, the mean square error of the random forest method is 0.813, average absolute error – 0.781, R-squared – 0.176. Accordingly, among the methods, the method of support vectors showed the best result by 6%. Among the signs, the signs that had the greatest influence on the prediction were determined to be the student's age, parents' education, free time and grades of the previous term.

In general, it can be concluded that the conducted research Machine learning algorithms allow to effectively analyze the signs that affect the quality of education of learners and to predict grades. We believe that it will be a useful tool for educational institutions and teachers in developing individual learning approaches and improving educational outcomes.

References

1. Lidia Sandra, Ford Lumbangaol, Tokuro Matsuo Machine Learning Algorithm to Predict Student's Performance: A Systematic Literature Review, Vol. 10, Issue 4, – 2021. – P. 1919-1927.

2. J. Dhilipan, N. Vijayalakshmi, S. Suriya, Arockiya Christopher Prediction of Students Performance Using Machine learning, – 2021 IOP Conf. Ser.: Mater. Sci. Eng. 1055 012122 DOI 10.1088/1757-899X/1055/1/012122

3. Hussein Altabrawee, Osama Ali, Amir Qaisar Predicting Students' Performance Using Machine Learning Techniques, Journal of University of Babylon for Pure and Applied Sciences 27(1):194-205 DOI:10.29196/jubpas.v27i1.2108

4. Y. Meier, J. Xu, O. Atan and M. Van Der Schaar, "Predicting grades," IEEE Trans. Signal Process, – 2016. – Vol. 64, – P. 959-972.

5. P. Guleria, N. Thakur, and M. Sood, "Predicting student performance using decision tree classifiers and information gain," Proc. 2014 3rd Int. Conf. Parallel, Distrib. Grid Comput. PDGC, – 2015. – P. 126-129.

6. Yıldız, M.B., Börekçi, C. (2020). Predicting Academic Achievement with Machine Learning Algorithms. Journal of Educational Technology & Online Learning, 3(3), 372-392.

7. P. Kaur et al. A review of machine learning techniques for student performance prediction. Conference: 2021 International Conference on Computer & Information Sciences (ICCOINS) – 2018. – 7p.

8. M. Al-Radaideh et al. Predicting Student Academic Performance Using Machine Learning Algorithms. –TEM Journal. Volume 10, Issue 4, Pages 1919-1927, ISSN 2217-8309, DOI: 10.18421/TEM104-56

9. J. M. Banda et al. Using Machine Learning Algorithms to Predict Final Grades of Students. Bakolori Journal of General Studies Vol. 12 No. 2 – 2019.

10. A. K. Sharma and A. K. Chaturvedi. A Comparative Study of Machine Learning Algorithms for Predicting Student Performance. International Journal of Intelligent Systems and Applications 11(12):34-45 DOI:10.5815/ijisa.2019.12.04

11. A. Mittal and N. Jindal. An Analysis of Machine Learning Techniques for Student Performance Prediction. *Educ. Sci.* 2021, *11*(9), 552; https://doi.org/10.3390/educsci11090552.

12. Al-Alawi, L., Al Shaqsi, J., Tarhini, A. et al. Using machine learning to predict factors affecting academic performance: the case of college students on academic probation. Educ Inf Technol (2023). https://doi.org/10.1007/s10639-023-11700-0

13. M. Abdollahi and S. K. Abdullah. Comparison of Machine Learning Algorithms for Student Performance Prediction in Higher Education. International Journal of Computer Sciences And Engineering 7(4):721-725

14. V. Nair et al. Prediction of Student Performance Using Machine Learning Algorithms: A review. – Proceedings of the International Conference on Applications of Machine Intelligence and Data Analytics (ICAMIDA 2022) (pp.735-741)

15. Albreiki, B.; Zaki, N.; Alashwal, H. A Systematic Literature Review of Student' Performance Prediction Using Machine Learning Techniques. Educ. Sci. 2021, 11, 552. https://doi.org/ 10.3390/educsci11090552

16. R. Rasheed et al. Predicting Student Performance in Higher Education using Machine Learning Algorithms. Journal of University of Babylon, Pure and Applied Sciences, Vol.(27), No.(1): 2019.

17. S. N. Tiwari and A. K. Misra. Predicting Student Performance in MOOCs using Machine Learning Techniques. – Conference: 11th International Conference on Data Mining, Computers, Communication and Industrial Applications (DMCCIA-2017) Kuala Lumpur (Malaysia)

18. Alyahyan, Eyman & Dustegor, Dilek. (2020). Predicting Academic Success in Higher Education Literature Review and Best Practices. International Journal of Educational Technology in Higher Education. 17. 10.1186/s41239-020-0177-7.