Enhancing Academic Success: Performance Early Prediction Using Machine Learning Algorithms

Abderrazek Hachani, Maha Mallek and Yosra Jmal

May 21, 2024

Enhancing Academic Success: Performance Early Prediction Using Machine Learning Algorithms

Abderrazek Hachani [1], Maha Mallek[1,2] and Yosra Jmal[1,3]

[1] ESPRIT School of Engineering
[2] LARIA, National School of Computer Science (ENSI)
[3]LTSIRS, National Institute of Applied Science and Technology (INSAT)

Abderrazek.hachani@esprit.tn
Maha.mallek@esprit.tn
Yosra.Jmal@esprit.tn

Abstract. Emerging technologies, particularly artificial intelligence (AI) and machine learning (ML) algorithms, present valuable opportunities to analyse learning management system data (LMS) and considered as the corner stone of Learning analytics (LA). The aim is to analyse student performance during a course or a whole academic year. In particular, it identifies at risk students and enables educators to timely support this student’s category and can provide clear guidance to improve teaching and learning strategies. This is why implementing an Early-Warning System is very crucial in this context to alert at risk student with weak performance, during first course sessions, to mitigate potential failures. The primary objective of this paper is to develop and implement an early-warning system that assists educators in identifying these students requiring attention and prompts them to be aware of their academic progress, thereby facilitating timely interventions to reduce the risk of failure. This study target web technologies course for the third-year engineering level in Esprit school of engineering, Tunisia. The experiments involve testing and comparing the performance of various classifiers, with a focus on Logistic Regression (LR), Random Forest (RF), Naive Bayes (NB) and K-Nearest Neighbours (KNN). The classification process considers factors such as students' engagement in different activities over time, the scores obtained in these activities, class attendance, and final results. The ultimate goal is to early predict student performance by categorizing them into two groups: those requiring additional support (convocation) and those who achieve well, the initial weeks of the course.

Keywords: Learning Management System (LMS), Machine Learning, Learning Analytics, Early Prediction, Academic Success, Educational Technology.

1 Introduction

When determining a student's knowledge, skills, and talents, learning evaluation in education entails a methodical procedure of obtaining and analyzing information. Accreditation programs and standards offer structures to guarantee uniformity and quality in education. Institutions can uphold high standards and show accountability by using these frameworks.

The measure of the extent to which each student achieves the intended specified learning outcomes, if we value personal and interpersonal skills, and product, process, system, and service building skills, and incorporate them into curriculum and learning experiences, then we must have effective assessment processes for measuring them. Different categories of learning outcomes require different assessment methods. For example, learning outcomes related to disciplinary knowledge may be assessed with oral, online and written tests, while those related to design-implement skills may be better measured with recorded observations. Using a variety of assessment methods accommodates a broader range of learning and increases the reliability and validity of the assessment data.

The future of learning experiences in the ever-changing field of education can be greatly shaped by the efficient use of data. Learning Management Systems (LMS) have emerged as crucial instruments for gathering enormous volumes of data about students, offering a previously unheard-of chance to learn more about the factors that predict academic success. The abundance of data created by learning management system (LMS) platforms has created opportunities for utilizing machine learning approaches to predict and comprehend student progress as educational institutions move more and more towards digital platforms.

This article presents a novel project at esprit school of engineering that uses data from Google classroom to forecast student performance by utilizing machine learning techniques. This project is driven by the goal of providing educators with proactive insights that will allow them to recognize students who might need more assistance and adjust their teaching methods to meet their specific needs. We expect to explore the complex patterns hidden in LMS data, which goes beyond conventional evaluation techniques.

The remainder of this paper is organized as follows: Related work on early-warning systems is presented in the Section II. Section III presents the research methodology. Using our purpose-built dataset, the experimental results of our system are presented in section IV. Finally, Section V concludes this paper and outlines future work.

2 RELATED WORK

There is a large body of literature relevant to explore machine learning techniques for predicting students' performance. In this section, we review the most recent and accurate works dealing with this problem.

Llanos et al. [1] have proposed a model for forecasting student success in a 16-week CS1 programming course. Grades, delivery time and the total number of attempts in exams are all used by the model. The model was trained and assessed using 8 algorithms and Week three saw the best results for the gradient boosting classifier.

In order to identify students who are at risk of failing in blended learning environments, the authors of Fahd et al. [2] suggests a novel strategy that makes use of machine learning models. According to the research, random forest algorithm obtained an accuracy of 85%.

The prediction of student performance by data mining and artificial intelligence is covered in paper [3] 16 features from 480 students make up the dataset extracted from Kalboard LMS, and the trained model's accuracy was 0.76.

In[4] the authors have asserted that Business Understanding, Data Understanding, Data Preprocessing, and Modelling constituted the four stages of the study project that were conducted. Decision tree, Bayesian, and k-Nearest Neighbor classifiers were the types of classifiers employed in the testing process, with the accuracy of predictions ranging between 52% and 67%.

The effectiveness of ML and LSTM-based models in forecasting student performance is assessed in paper [5] The study used the LIME method for interpretability analysis and focused on online teaching and learning using a virtual learning environment. The authors' main conclusions showed that deep learning techniques perform better at predicting grades than traditional regression techniques. Conversely, interpretability decreases with increasing prediction model sophistication.

To evaluate students' success in the course, Zangooei et al. [6] have employed learning analytics technologies. The LSTM neural network model was utilised to forecast pupils' academic achievement. In terms of prediction accuracy, the authors demonstrated that LSTM network performs better than the SVM method.

3 RESEARCH METHODOLOGY

As shown in fig.1, the research methodology for this study follows a structured approach to predict early risk students based on Google Classroom data. The methodology encompasses several key stages, beginning with the clear definition of the research objective. A comprehensive review of existing literature on student risk prediction, machine learning, and educational data mining was conducted to establish a foundation for the study.

Data collection involved obtaining LMS data, ensuring compliance with data privacy regulations and ethical considerations. The collected data underwent rigorous preparation, including cleaning, handling missing values, and preprocessing tasks such as normalization and encoding.

Feature extraction was performed to identify and extract relevant features from the LMS data, which would serve as input for the prediction models. Subsequently, Pearson correlation analysis was conducted to assess relationships between variables and their correlation with early student risk.

The predictive models were built using Logistic Regression, Random Forest, Naïve Bayes and K-Nearest Neighbors. The data was split into training and testing sets, and the models were trained and evaluated using appropriate performance metrics such as accuracy.

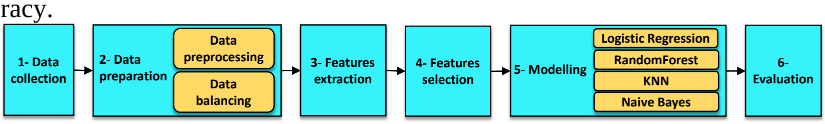

Fig. 1. Overview of the different steps in our proposed method.

3.1 Data Collection

The dataset utilised in this study consists of information derived from the web technologies course classroom for third-year engineering students conducted in the 2023 academic year at Esprit School of Engineering, Tunisia. This dataset includes details regarding students' assessments and Student Attendance.

The classroom setting facilitates flexible curriculum delivery, giving teachers the freedom to assign and review content with their pupils. With this flexibility, students can take advantage of opportunities to participate in online activities, access course materials, and take tests, all of which can improve their overall learning experience.

After analysis, a csv file is produced from the data collection. There are 120 rows and 15 columns of gathered records in this file. Students' names and identifiers are included in the columns along with the dates of submission for assessments 1, 2, 3, and 4. In a similar vein, all seven class periods' worth of student attendance is included. The overall average and the final exam mark for this module come next. Every row represents a student record. To train, validate, and test the prediction models, these data were used. A total of 120 students took the course; 24 students, or 80% of the data, were set aside for testing, while the remaining 96 students, or 20% of the data, were randomly assigned for training and validation.

3.2 Data preparation

Any data mining method must start with data preparation. This is the initial stage of data preparation for early student performance prediction. Two primary tasks are involved: (i) **Data anonymization** that aim to make the data anonymous, the two columns that contained the names and identities of the students were deleted. **(ii) Data pre-processing** that first use df.isna(). sum() to find the missing values in the DataFrame. The.dropna() function was then used to eliminate these records. The following stage was to convert the results of the four evaluations' submission dates into a numerical representation according to how those results related to a target date. This was accomplished by using the convert_to_float function. Following was the encoding of the assessment results: There are four categories for student submissions: three for those turned in before the deadline, two for those sent in on time, one for those turned in after the deadline, and zero for those that are missing. Furthermore, the attendance of students was converted into a percentage for every individual student. The last stage of data preprocessing was creating values that represented success and failure from the student's final exam and total grade columns.

3.3 Features extraction

Activities conducted by students within the Learning Management System (LMS) classroom involve diverse tasks, including the completion of four homework assignments with specified deadlines. To capture this information, the first feature extracted is the submission date for each student's assessment in the classroom. Additionally, we incorporated other pertinent features such as student attendance across the seven sessions, the grade obtained in the final exam for the module, and the overall grade achieved. This analysis revealed the identification of four key features: submission date for each assessment, final grade, student attendance during the seven sessions, and overall grade.

3.4 Features selection

Selecting features with correlations close to zero. The Pearson method was employed to calculate the linear correlation between pairs of features, producing results between -1 and +1. Upon conducting these calculations, it was observed that the overall grade feature exhibited a correlation coefficient of 1 with other features. Consequently, the overall average was eliminated as it does not contribute significantly.

3.5 Modelling

Modelling is a fundamental step of the presented method which follows the preprocessing of the dataset and the features selection. This task requires two steps: **(i) Selected algorithms***:* The suggested model incorporates four algorithms: K-Nearest Neighbours (KNN) [7] Random Forest (RF) [8], Logistic Regression (LR)[9] and Naive Bayes (NB)[10][11]. The selection of RF, LR, and NB algorithms is based on their ability to accurately forecast student performance in the initial phases, as suggested by baseline articles. Furthermore, this choice is driven by the unique benefits that each algorithm provides in this situation. The construction of the prediction model comes after the algorithm selection. In order to do this, the model is trained using 80% of the data produced during the data preparation step, with the remaining 20% set aside for testing. The final predictions are then produced using the selected features, and the results for specific metrics during weeks 2, 4, and 7 of the courses are obtained.
(ii) Selected metrics*:* In order to evaluate the performance of the used algorithms, we adopt the official evaluation metric, which is based on F1-score. The F1 score can be interpreted as a weighted average of the precision and recall.

3.6 Evaluation

The goal of this research is to forecast student performance beginning in week two of the course by achieving an F1 score metric value of more than 70%.

4 RESULTS

In this study, a total of four traditional classification algorithms were utilized for early prediction student performance: Naive Bayes, Logistic Regression (LR), k-Nearest Neighbors (KNN) and Random Forest (RF). Various tests were conducted to determine the performance of these algorithms using F1 score. This metric is evaluated for weeks 2, 4, and 7 of the study as shown in table 1.

Across the weeks, the algorithm with the highest performance result was the LR. The LR predicted the student performance with F1 score 72.56% in the week 2, while 76.37% in the week 4 and 84.61% in the seventh week. It can be remarked, that F1 score increases during the weeks and the loss values decrement with increasing weeks, which indicates the strength of the model. Conversely, KNN and RF achieved the lowest results. These models could be a potentially good method, but it need a much higher number of students to produce better and relevant results.

Table 1. Results of Prediction algorithms in Weeks 2, 4, and 7.

Week	Metrics	KNN (%)	LR (%)	RF (%)	NB (%)
Week2	F1-Score	47.30	**72.56**	63.37	47.30
Week4		56.09	**76.37**	68.91	56.09
Week7		59.80	**84.61**	71.64	59.80

5 CONCLUSION

In conclusion, this study has demonstrated the potential of machine learning algorithms in predicting early risk students based on Learning Management System (LMS) data. The research methodology employed a systematic approach, encompassing data collection, preparation, feature extraction, correlation analysis, and the implementation of predictive models using Logistic Regression, Random Forest, and Naïve Bayes algorithms.

Ethical concerns about student data protection and appropriate use were considered in this work. The study of the implementation outcomes highlighted performance of each classification algorithm and provided insightful information about how well it performed. Despite the short dataset in this study, the Logistic Regression (LR) approach outperforms Random Forest (RF), k-Nearest Neighbours (KNN), and Naive Bayes.

We are thinking about including new aspects in this study, like student psychologies, personal data, and additional weekly activities, to improve it. We'll also broaden the study to include early risk pupils in a variety of academic programmes in addition to a particular class.

6 REFERENCES

1. Llanos, J., Bucheli, V. A., & Restrepo-Calle, F.: Early prediction of student performance in CS1 programming courses. PeerJ Computer Science, 9, e1655. (2023).
2. Fahd, K., Miah, S. J., & Ahmed, K.: Predicting student performance in a blended learning environment using learning management system interaction data. Applied Computing and Informatics. (2021).
3. Bhusal, A.: Predicting Student's Performance Through Data Mining. arXiv preprint arXiv:2112.01247 (2021).
4. López Zambrano, J., Lara Torralbo, J. A., & Romero Morales, C. : Early prediction of student learning performance through data mining: A systematic review. Psicothema. (2021).
5. Chen, H. C., Prasetyo, E., Tseng, S. S., Putra, K. T., Kusumawardani, S. S., & Weng, C. E.: Week-Wise Student Performance Early Prediction in Virtual Learning Environment Using a Deep Explainable Artificial Intelligence. Applied Sciences, 12(4), 1885.(2022).
6. Zangooei, H., & Fatemi, O.: Predicting students at risk of academic failure using learning analytics in the learning management system. Quarterly of Iranian Distance Education Journal, 3(2), 32-44. (2021).
7. Syed, M. E.: Attribute weighting in k-nearest neighbor classification (master's thesis). (2014).
8. Siemens, G.: Learning analytics: envisioning a research discipline and a domain of practice. In Proceedings of the 2nd international conference on learning analytics and knowledge (pp. 4-8). (2012, April).
9. Hosmer, D. W., Hosmer, T., Le Cessie, S., & Lemeshow, S. : A comparison of goodness-of-fit tests for the logistic regression model. Statistics in medicine, 16(9), 965-980. (1997).
10. Domingos, P., & Pazzani, M.: On the optimality of the simple Bayesian classifier under zero-one loss. Machine learning, 29, 103-130. (1997).
11. Levine, R. R.: Factors affecting gastrointestinal absorption of drugs. The American journal of digestive diseases, 15, 171-188.(1970).