https://doi.org/10.36315/2023v1end047
Education and New Developments 2023

EARLY IDENTIFICATION OF ACADEMIC FAILURE ON HIGHER EDUCATION: PREDICTING STUDENTS' PERFORMANCE USING AI

Fidel Cacheda, Manuel F. López-Vizcaíno, Diego Fernández, & Víctor Carneiro
Center for Information and Communications Technologies Research (CITIC)
Department of Computer Science and Information Technologies, Campus de Elviña s/n, 15071 (Spain)

Abstract

In this work we focus on the early identification of academic failure in higher education as a mean to allow educators to provide an early intervention and help students on a risky position to achieve academic success. For this purpose, we define a dataset of more than one thousand students with their respective grades collected from a Computer Networks course on a Computer Science degree at a Spanish university throughout four years. From the dataset we extract different features corresponding to the laboratory and quiz assignments proposed to the students during the course that intend to represent the effort and accomplishment achieved by the students. A preliminary analysis of the dataset shows a potential relation between the scores achieved throughout the course and the final exam mark. The aim is to predict if a student will pass or not the final exam using only information extracted from the different laboratory and quiz assignments. In this sense, we define a data mining classification task following a supervised learning approach where a selection of well-known machine learning algorithms is evaluated following a 10-fold cross-validation scheme to assess the performance and robustness of the models. Our results show that using Random Forest we can accurately predict in more than 91% of the cases if a student will pass or not the final exam, achieving a F1-score of 0.916. Moreover, we perform a feature importance analysis highlighting how laboratory assignments features have a higher contribution to the learning model than quiz assignments.

Keywords: *Early identification, higher education, academic failure, machine learning, artificial intelligence.*

1. Introduction

Student success has been defined by York et al. as "academic achievement, satisfaction, acquisition of skills and competencies, persistence, attainment of learning objectives, and career success" (York, Gibson, & Rankin, 2015). In fact, student success is considered a key metric on higher education institutions for assessing their quality (Alyahyan & Dustegor, 2020).

The use of Artificial Intelligence and, more specifically, Data Mining techniques allow us to mine large amounts of data and education is one important field where Data Mining can be applied. In fact, Educational Data Mining (EDM) has risen as a research field that involves statistics, data mining and machine learning, and other fields to analyze educational big data effectively (Xiao, Ji, & Hu, 2022), (Batool et al., 2022).

In this article, we focus on the early identification of academic failure in higher education as a mean to allow educators to provide an early intervention and help students on a risky position to achieve academic success. A dataset is defined with more than one thousand students' grades on a Computer Networks subject. Several features corresponding to the laboratory and quiz assignments are extracted and a data mining classification task is proposed to predict if a student will pass or not the final exam.

2. Related works

EDM comprehends multiple research works involving the discovery of knowledge patterns about educational facts and the learning process (Anoopkumar & Rahman, 2016), such as performance (Saa, 2016), success (Martins, Miguéis, Fonseca, & Alves, 2019), satisfaction (Alqurashi, 2019) or dropout rate (Pérez, Castellanos, & Correal, 2018), among others.

Our work is more related with the prediction of students' academic performance. In this sense, the authors of (Mueen, Zafar, & Manzoor, 2016) test three classification algorithms (Naïve Bayes, Neural Network, and Decision Trees) to predict students' performance on two undergraduate courses. Sivasakthi applies different classification algorithms to predict programming performance on a Computer Application course proposing a knowledge flow model (Sivasakthi, 2017). In (Putpuek, Rojanaprasert, Atchariyachanvanich, & Thamrongthanyawong, 2018) the students' performance is predicted based on their personal background, including gender, scholarship awarded, previous educational background, admission type, talent and province of high school, although a moderated accuracy was achieved. In (Almarabeh, 2017) a simple comparison of different classification algorithms is presented using a dataset of 225 students. Yassein et al. in (Yassein, Helali, Mohomad, et al., 2017) search for patterns to predict students' performance and discover that the most affecting factor is class attendance. More recently, (Alsariera et al., 2022) analyses some research works published between 2015 and 2021, concluding that machine learning can be beneficial to identify various academic performance areas.

These previous works are related with ours in the sense that the aim is to predict students' academic performance, although the early detection of a potential low performance is also relevant in our research.

3. Course description

This work has been performed collecting the data from a subject on Computer Networks taught at the degree in Computer Science Engineering at the University of A Coruña (Spain). This subject is taught in the second semester of the second year and it takes 6 credits of European Credit Transfer System (ECTS), which correspond to 60 hours of classroom teaching plus 90 hours of personal work.

The course is focused on the main aspects of networking, including the main features, functionalities and structure of computer networks and Internet. This subject constitutes the first approach to computer networking for most students and the main objective is that students understand the different layers and protocols that come into action when two devices communicate using TCP/IP.

The subject has assigned four sessions per week (one hour per session): two theoretical sessions on different days and two consecutive sessions for the laboratory. The syllabus of the course is as follows:

- Topic I – Introduction to computer networks, Internet and TCP/IP
- Topic II – Application layer: Web, email and DNS
- Topic III – Transport layer: UDP and TCP
- Topic IV – Network layer: IP, subnetting and routing
- Topic V – Link layer: ARP, Ethernet and WiFi

Throughout the course, each student must individually develop and present the following laboratory projects, which are not mandatory:

- Project I: Introduction to socket programming in Java
- Project II: Basic Java Web server
- Project III: Introduction to Network simulation with Cisco Packet Tracer
- Project IV: Network simulation – Subnetting and routing

Also, students are presented with two quizzes throughout the course that must solve online. These quizzes are composed of questions from the theory lessons and are intended to reinforce the students' continuous learning. The first quiz covers topics I and II, while the second quiz covers topics III and IV.

The evaluation of the subject includes a theoretical exam (two calls are available for the students, one at the end of the semester and another approximately one month later) that corresponds to 70% of the final grade and the students are required to achieve at least a grade of 4 (out of 10) to compute the final grade. The final grade also includes the laboratory and quiz grades, as 25% and 5%, respectively (no minimum grade is required in this case). To pass the subject, a final mark greater than or equal to 5 must be achieved by the student.

4. Dataset and features

4.1. Dataset

We have built a dataset collecting the grades from the Computer Network subject presented on the previous section throughout four years: 2017-18, 2018-19, 2020-21, 2021-22. We skip the year 2019-20 because the Covid pandemic produced changes on the students' evaluation.

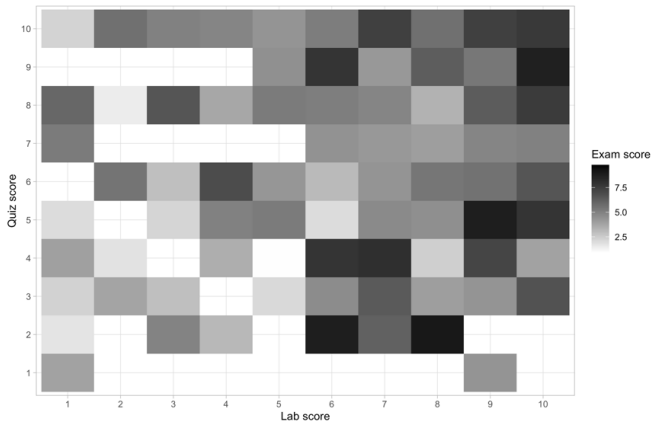
Table 1. Dataset summary.

	Number of students	Laboratory		Quiz		Exam	
		Pass	Fail	Pass	Fail	Pass	Fail
2021-22	256	134	122	177	79	169	87
2020-21	289	170	119	225	64	197	92
2018-19	244	169	75	159	85	188	56
2017-18	261	181	80	175	86	189	72
Total	**1050**	**654**	**396**	**736**	**314**	**743**	**307**

Table 1 presents a summary of the main characteristics of the dataset. The dataset is composed of more than one thousand students with their respective grades. We summarize the number of students that pass and fail each one of the main evaluation parts. We consider that a student passed the exam if in any of the two calls she/he achieved a score higher than or equal to 4. Also, the number of students that failed the exam includes the students that did not show up. For evaluation purposes, only students that actually did the exam will be taken into consideration which reduces the total number of failed exams to 158.

Our intuition is that students that perform well during the course (i.e. in the laboratory and quizzes) will tend to also perform well in the final exam. Figure 1 presents a heat map of the exam scores with respect to the laboratory (X axis) and quiz (Y axis) scores. From the figure we can observe how darker tones (corresponding to higher exam scores) are located on the right half of the figure and, specially, on the upper corner corresponding to higher grades on both laboratory and quiz assignments, confirming our intuition.

Figure 1. Heat map of exam scores with respect to laboratory and quiz scores.



4.2. Features

From the dataset, we extract several features that are divided into two groups, depending on if they correspond to laboratory or quiz grades. All scores are normalized to operate between 0 and 1.

Laboratory features include the following:

- Laboratory assignments scores: one feature for each assignment (denoted as Lab1, Lab2, Lab3 and Lab4).
- Laboratory score (Lab_score): final laboratory score.
- Laboratory passed (Lab_passed): boolean value indicating if the student passed the laboratory assignments (i.e. lab score higher than or equal to 5).
- Laboratory effort (Lab_effort): percentage of assignments submitted.
- Average, standard deviation and median for laboratory assignments scores (Lab_avg, Lab_std and Lab_median)
- Number of passed laboratory assignments (N_lab_passed)
- Number of laboratory assignments submitted (N_lab_tried)

For the quizzes, analogous features have been extracted.

Moreover, the aggregation of laboratory and quiz scores was calculated as the average (denoted as LabQuiz_score).

5. Data mining problem

We focus on one task: to predict if a student will pass the final exam, just taking into consideration the work done by the student throughout the course in terms of laboratory and quiz assignments.

p-ISSN: 2184-044X e-ISSN: 2184-1489 ISBN: 978-989-35106-3-6 © 2023

For this purpose, we define a data mining classification task following a supervised learning approach. We consider the following standard off-the-shelf machine learning algorithms that intend to cover the main techniques: J48, JRip, LibLinear, Logistic Regression (LR), Naïve-Bayes (NB), Random Forest (RF) and SVM.

The evaluation is conducted following a 10-fold cross-validation scheme to assess the performance and robustness of the models. To address the class imbalance (743 students passed the exam vs. 158 that failed) we oversample the minority class using Synthetic Minority Oversampling Technique (SMOTE). As evaluation metrics, we report our results on the percentage of accurately predicted instances, F1-measure, Receiver Operating Characteristics (ROC) Area Under the Curve (AUC), Precision Recall Curve (PRC) AUC and Root Mean Squared Error (RMSE).

6. Experimental results

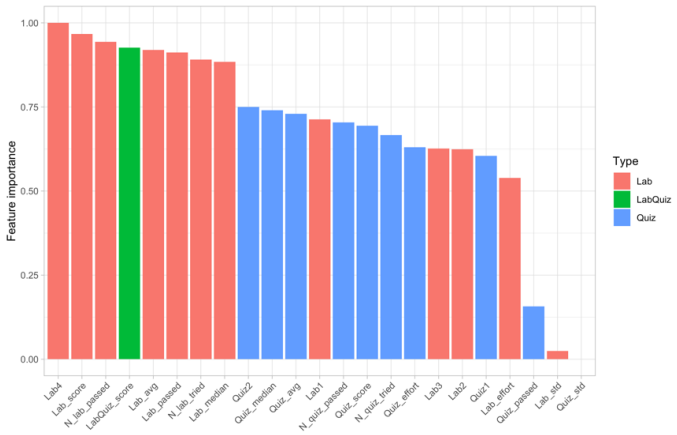
In Table 2, we present the results for the models trained on the proposed dataset. Random Forest is the best performing model, being able to predict accurately more than 91% of the cases and the F-score is 0.916. Also, for all remaining metrics, RF is consistently the best performing model.

Table 2. Results for failure detection using all features.

Model	Correctly classified	F1	ROC AUC	PRC AUC	RMSE
J48	88.12%	0.881	0.866	0.795	0.330
JRip	87.06%	0.871	0.873	0.830	0.332
LibLinear	85.20%	0.852	0.825	0.737	0.384
LR	84.80%	0.848	0.890	0.878	0.338
NB	77.84%	0.777	0.837	0.777	0.459
RF	**91.64%**	**0.916**	**0.939**	**0.897**	**0.259**
SVM	82.42%	0.824	0.799	0.707	0.419

We perform an ablation study, repeating the evaluation considering only laboratory features and only quiz features. In both cases, results did not improve the best performing model from Table 2. In general terms, using only laboratory features achieved better results than using only quiz features. This result is expected, since the laboratory assignments must be developed individually by each student, while quiz assignments can be answered collaboratively and, therefore, may not reflect accurately the student effort and knowledge in the subject.

Figure 2. Features importance.



Finally, we analyze feature importance on Figure 2 by measuring Pearson's correlation between each feature and the class. We applied min-max normalization to Pearson's correlation values obtained.

Laboratory features are represented in red, while quiz features are showed in blue. The feature aggregating both values is displayed in green. From the figure, we can observe how laboratory features are more important for the classification task than quiz features confirming the results from the ablation study. Also interesting is the high position in the ranking for the aggregation feature LabQuiz_score.

7. Conclusions

In this work we have showed how following a supervised learning approach and using only information extracted from the grades obtained in laboratory and quiz assignments, we are able to predict if a student will pass or final the final exam in more than 91% of the cases. Moreover, our feature performance analysis shows how laboratory assignments features have a higher contribution to the learning model than quiz assignments.

In the near future, we expect to apply these results throughout the course to identify students on a risky position that may require further supervision and evaluate their potential improvement.

Acknowledgements

This research was supported by the Ministry of Economy and Competitiveness of Spain and FEDER funds of the European Union (Project PID2019-111388GB-I00) and by the Centro de Investigación de Galicia "CITIC", funded by Xunta de Galicia and the European Union (European Regional Development Fund- Galicia 2014-2020 Program), by grant ED431G 2019/01.

References

Almarabeh, H. (2017). Analysis of students' performance by using different data mining classifiers. *International Journal of Modern Education and Computer Science*, *9*(8), 9.

Alqurashi, E. (2019). Predicting student satisfaction and perceived learning within online learning environments. *Distance Education*, *40*(1), 133– 148.

Alsariera, Y. A., Baashar, Y., Alkawsi, G., Mustafa, A., Alkahtani, A. A., & Ali, N. (2022). Assessment and evaluation of different machine learning algorithms for predicting student performance. *Computational Intelligence and Neuroscience*, 2022.

Alyahyan, E., & Dustegor, D. (2020). Predicting academic success in higher education: literature review and best practices. Internat. *Journal of Educational Technology in Higher Education*, *17*(1), 1–21.

Anoopkumar, M., & Rahman, A. M. Z. (2016). A review on data mining techniques and factors used in educational data mining to predict stu- dent amelioration. In 2016 international conference on data mining and advanced computing (sapience) (pp. 122–133).

Batool, S., Rashid, J., Nisar, M. W., Kim, J., Kwon, H.-Y., & Hussain, A. (2022). Educational data mining to predict students' academic perfor- mance: A survey study. *Education and Information Technologies*, 1–67.

Martins, M. P., Miguéis, V. L., Fonseca, D., & Alves, A. (2019). A data mining approach for predicting academic success–a case study. In International conference on information technology & systems (pp. 45–56).

Mueen, A., Zafar, B., & Manzoor, U. (2016). Modeling and predicting students' academic performance using data mining techniques. *Internat. journal of modern education & computer science*, 8(11).

Pérez, B., Castellanos, C., & Correal, D. (2018). Predicting student drop- out rates using data mining techniques: A case study. In Ieee colombian conference on applications in computational intelligence (pp. 111–125).

Putpuek, N., Rojanaprasert, N., Atchariyachanvanich, K., & Thamrongthanya- wong, T. (2018). Comparative study of prediction models for final gpa score: a case study of rajabhat rajanagarindra university. In 2018 ieee/acis 17th international conference on computer and information science (icis) (pp. 92–97).

Saa, A. A. (2016). Educational data mining & students' performance prediction. *International Journal of Advanced Computer Science and Applications*, *7*(5).

Sivasakthi, M. (2017). Classification and prediction based data mining algo- rithms to predict students' introductory programming performance. In 2017 international conference on inventive computing and informatics (icici) (pp. 346–350).

Xiao, W., Ji, P., & Hu, J. (2022). A survey on educational data mining methods used for predicting students' performance. *Engineering Reports*, *4*(5), e12482.

Yassein, N. A., Helali, R. G. M., Mohomad, S. B., et al. (2017). Predicting stu- dent academic performance in ksa using data mining techniques. *Journal of Information Technology & Software Engineering*, 7(5), 1–5.

York, T. T., Gibson, C., & Rankin, S. (2015). Defining and measuring academic success. *Practical assessment, research, and evaluation*, *20*(1), 5.