Expert Systems

WILEY

Expert Systems

ORIGINAL ARTICLE OPEN ACCESS

Unveiling the Differences in Early Performance Prediction Between Online Social Sciences and STEM Courses Using Educational Data Mining

Raúl Marticorena-Sánchez | Antonio Canepa-Oneto | Carlos López-Nozal | José A. Barbero-Aparicio

Departamento de Ingeniería Informática, Universidad de Burgos, Burgos, Spain

Correspondence: Raúl Marticorena-Sánchez (rmartico@ubu.es)

Received: 17 May 2024 | **Revised:** 11 December 2024 | **Accepted:** 2 January 2025

Funding: The authors received no specific funding for this work.

Keywords: educational data mining | learning analytics | learning at scale | online learning logs | student performance prediction | supervised data-mining

ABSTRACT

Educational Data Mining and Learning Analytics in virtual environments can be used to diagnose student performance problems at an early stage. Information that is useful for guiding the decisions of teachers managing academic training, so that students can successfully complete their course. However, student interaction patterns may vary depending on the knowledge domain. Our aim is to design a framework applicable to online Social Sciences and STEM courses, recommending methods for building accurate early performance prediction models. A large-scale comparative study of the accuracy of multiple classifiers applied to classify the interaction logs of 32,593 students from 9 Social Sciences and 13 STEM courses is presented. Corroborating the results of other works, it was observed that high early performance prediction accuracy was obtained based on nothing other than student logs: accuracies of 0.75 in the 10th week, 0.80 in the 20th week, 0.85 in the 30th week and 0.90 in the 40th week. However, accuracy rates were observed to vary significantly, in relation to the classification algorithm and the knowledge domain (Social Sciences vs. STEM). These predictions are generally less accurate for Social Sciences compared to STEM courses, especially at the beginning of the course, with fewer differences observed in the final weeks. Additionally, this research identifies instances of low-accuracy outliers in the prediction of Social Sciences courses over time. These findings highlight the complex challenges and variations in early performance prediction across different domains in online education.

1 | Introduction

Stakeholders with a role in the management of online courses that generate thousands of student interaction logs stored in Virtual Learning Environments (VLE) need specialised Educational Data Mining (EDM) and Learning Analytics (LA) tools (Peña-Ayala 2014; Romero and Ventura 2020; Romero, Ventura, and García 2008). Both, EDM and LA are recognised as fundamental components for the management of learning improvement actions that are embedded within Higher Education (HE) learning processes. Moreover, studies on the application of learning improvement actions, some examples of which can be seen in Aldowah, Al-Samarraie, and Fauzy (2019) and in other previous works (Peña-Ayala 2014; Romero, Ventura, and García 2008), are growing.

One fundamental task involved in learning improvement actions is student performance prediction (Rastrollo-Guerrero, Gomez-Pulido, and Domínguez 2020). Several supervised machine-learning techniques have been applied to predict

learning outcomes: (i) identifying students at "high risk" of dropping out and (ii) predicting the future performance of students and their final exam grades (Tomasevic, Gvozdenovic, and Vranes 2020).

In the particular context of the Massive Open Online Course (MOOC), the success rate is very low, reaching an average value of 47.20%, according to Kuzilek, Hlosta, and Zdrahal (2017). Intelligent tutoring systems can incorporate student learning analytics with available VLE information. The data used in these systems can be from multiple sources: (i) student behaviour measured through learning object interaction (*a.k.a. logs*), (ii) intermediate learning activity grades (Waheed et al. 2020), and (iii) student demographics like age, gender, region, previous education, *etc*. (Rizvi, Rienties, and Khoja 2019).

Student interaction with online learning environments appears to generate sufficient data volumes for the analysis and the definition of prediction methods (Tomasevic, Gvozdenovic, and Vranes 2020). The temporal analysis of logged interactions is undoubtedly a promising pathway for predicting performance and identifying student behaviour, although it was noted in the review by Knight, Wise, and Chen (2017) that further investigation is needed within this area.

Moreover, research into student behaviour within virtual environments and the knowledge domains of Social Sciences (SS) and Science Technology Engineering, and also Mathematics (STEM) is not new. In Finnegan, Morris, and Lee (2008), significant differences between student online participation, persistence, and achievement across these domains were noted. In Sáiz-Manzanares et al. (2021), a study on small-scale VLE records (i.e., Moodle) noted some differences between the frequency of access to activities, taking into account gender and SS vs. STEM. However, to the best of our knowledge, such differences within the context of MOOC have not been confirmed in rigorous and large-scale experimental studies (Tomasevic, Gvozdenovic, and Vranes 2020). Thus, the availability of proper experimental and analytic design may enhance the accuracy of early–student failure prediction models and has yet to be considered in experimental research papers.

Large-scale learning studies that take place with multiple learners and high learner-to-facilitator ratios can yield useful results when generalising the data on logged interactions and learning experiences, one example being Roll, Russell, and Gaševic (2018). High volume, anonymised, open data sets, which meet both the ethical and the data privacy requirements of the participating institutions are therefore needed to establish an accurate experimental framework.

Following this line of educational research, the principal motivation of this study is to help stakeholders (i.e., institutional policymakers, faculty members, researchers and software designers) to guide the implementation of early–student failure detection and student success-rate prediction models on online courses. For this purpose, a large-scale, accurate, interpretable, and generalizable experimental framework was designed for VLE platforms, taking into consideration both the Social Sciences and the STEM domains.

The analysis reveals significant differences in prediction accuracy between Social Sciences and STEM course outcomes, with lower accuracy in the case of Social Sciences. At the beginning of the courses, the difference in prediction accuracy between Social Sciences and STEM is more pronounced, probably due to the structured nature of STEM content. In the later weeks, the differences in prediction accuracy between Social Sciences and STEM courses decrease as more data become available. Despite this, low-accuracy outliers remain more common in Social Sciences predictions, reflecting the variability and interpretive nature of the field.

The remainder of the paper is organised as follows. In Section 2, a literature review of related studies is summarised, paying particular attention to the use of student logs for the study of online performance prediction. In Section 3, the data set and the main concepts are described with which we quantify and compare the accuracy of student performance prediction algorithms. In Section 4, the results of measuring prediction algorithm accuracy, in accordance with various hypotheses concerning early prediction and course domain knowledge, are presented. In Section 5, the implications of the study and its possible applications are discussed, and finally, the main conclusions are summarised in Section 6.

2 | Related Work

The literature review presented in this section is structured into three subsections. In the first, Section 2.1, the literature on predicting student performance, particularly early performance and dropout prediction problems, is presented from a generic point of view. In the second, Section 2.2, large-scale student performance prediction analyses of both SS and STEM courses are presented. To do so, we extracted information from the Open University Learning Analytics Dataset (OULAD), a public data set (Kuzilek, Hlosta, and Zdrahal 2017). Since the compilation of the OULAD database, a great deal of research work has been conducted to predict student performance. However, there is still a highly varied set of approaches, methods and model designs that are constantly evolving and that offer different solutions to achieve the same objective. A state-of-the-art analysis of the data extracted from the OULAD data set will characterise the research work and quantify the research effort in certain areas. In the third subsection, Section 2.3, we identify the emerging trends and research gaps in the field of student performance prediction.

2.1 | Predicting Students' Performance

There is a large body of research work related to the task of predicting student final mark performance. Rastrollo-Guerrero, Gomez-Pulido, and Domínguez (2020) surveyed 70 papers on student online activities, including objectives, techniques, algorithms and methods. Each referenced work contained an analysis of data sets built from student activities at a particular level: school, high school and university. The objectives were connected to interests and risks associated with student-learning processes: student-dropout rates, student performance, recommended activities, resources and student knowledge (e.g.,

automated techniques that detect student learning styles (Karagiannis and Satratzemi 2018)). Different data-processing algorithms, methods and tools for the analysis of the above objectives are fundamental in these techniques for predictive purposes.

Alsariera et al. (2022) presented a review of 39 papers on approaches to Machine Learning (ML) and key student performance prediction features. Their results identified the six most widely used ML models: Decision Trees (DT), Artificial Neural Network (ANN), Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Linear Regression (LinR) and Naïve Bayes (NB). The study found that DT methods generated the models with the highest interpretability, making them more useful for designing effective student support (Helal et al. 2018).

In the scientific literature, many papers have been focused on student performance and dropout-rate prediction using small-scale data sets. A maximum data set size of 20,000 records in the work of Roy et al. (2018) was referenced in the systematic review of Alsariera et al. (2022). Another important conclusion of these review papers are their findings on the most relevant features used in the data sets, highlighting student interactions and grades. Academic and demographic data, internal assessment and family/personal attributes were other features also included in the data sets.

On the contrary, large-scale experimentation requires high-volume data sets with huge numbers of records (>10,000,000) that comply with anonymity and open data. Data sets such as OULAD (Kuzilek, Hlosta, and Zdrahal 2017) and HarvardX/MITX (Ho et al. 2014) meet these requirements. Al-Shabandar et al. (2019) defined a large-scale study using both of the aforementioned data sets. However, no account was taken of all the features in each data set in their study, because it unified the common features, selecting only general student interactions.

The model prediction accuracies of most experimental designs that use their particular data set vary between 78.0% (Rivas et al. 2021) and 98.2% (Alsariera et al. 2022). Comparative studies have shown that ANNs are the best performing method, although with no clearly significant differences (Alsariera et al. 2022). Many works create predictive models with the log files generated for the whole course, but those models are not useful for early prediction, because the log records used for prediction purposes are not used to train the models. ML applied to the early stages of a course for the prediction of student outcomes is a far less-well explored field of investigation. Gray and Perkins (2019) focused on attendance features in face-to-face university classes, predicting early failure with high accuracy (86%) in the 4th week of a 30 week course. Riestra-González, del Puerto Paule-Ruíz, and Ortin (2021) defined an early prediction model that rather than dependent on particular course features, only processed VLE records. They predicted student performance at 10%, 25%, 33% and 50% of the course delivery. The ANN yielded the best performance, with an accuracy of 80.1% after students had completed 10% of the course and 90.1% having reached the halfway mark.

Helal et al. (2018) predicted academic performance on the basis of student heterogeneity. Their model incorporated data from student enrollment forms, such as socio-economic and attendance type (e.g., full-time vs. part-time). The experimental results validated the hypothesis that the models trained with the instances forming the student sub-populations outperformed the models comprising all the data instances. The results also revealed that enrollment and course activity features provided useful information for identifying vulnerable students with higher precision. In some studies, the relationship of STEM studies with socio-economic factors of students, such as the gender gap, were analysed from the point of view of heterogeneity in student sub-populations. Wang and Degol (2017) proposed evidence-based recommendations for policy and practice, to improve STEM diversity and suggestions for future research directions.

2.2 | OULAD Research Papers

OULAD stores student-learning features together with their VLE interaction data, their intermediate and final grades, and demographic data. Designed for large-scale experimentation, OULAD provides access to 10,655,280 logs of 32,593 students, enrolled on 22 courses. Its large-scale characteristics facilitate the generalisation of experimental results. In addition, all interactions and grades are time stamped, so experiments can be designed that address early prediction.

We reviewed 14 research papers that featured experimental work with the OULAD dataset. Three data settings were considered for classifying these papers. The first, "Scale", informed us whether the combined data of all the courses had been used in the study or whether they were grouped by course. In second place, "Features" reflected the information used in the student prediction models (i.e., demographic, interaction and performance). The third one, "Early detection/period", represented the time component and whether it was used to model the predictions. Concerning the design used to define the prediction models, we considered another three characteristics. The first was the prediction "Method", which can take the following values: baseline (b), ensembles (e), neural networks (nn) and time series (ts). The second was "Technique", informing us whether a classifier (c) was applied in the research work; and the number of classes to predict (i.e., pass and fail), and whether a regressor (r) was used. We called the third one "Model", which informed us whether a single generic model (s) was created or whether multiple customised models organised by courses (c) and/or by time periods (wdt) were implemented. In the remainder of this section, we will detail these features and their application to the OULAD data set in the 13 above-mentioned papers (see Table 1).

Two large-scale experimental approaches were observed. The first one was adopted to generalise as much as possible, by using all interaction data from all the courses together (Adnan et al. 2022; Haiyang et al. 2018; Hassan et al. 2019; Waheed et al. 2020). With this approach, course-specific information was pre-processed to unify course contents (e.g., to aggregate all types of learning activities). The second approach was aimed at specialising course interaction, with the objective of maximising the performance of the models that were generated with data from the course (Al-Shabandar et al. 2019; Peach et al. 2019; Qiao and Hu 2020). Whether all the data or only data from some courses were used in the research is indicated under the column heading "Scale", in Table 1.

TABLE 1 | Summary of OULAD research papers on model-building characteristics.

	Data settings			Experimental design		
Cite	**Scale**	**Feature**	**Early/period**	**Method**	**Technique**	**Model**
Adnan et al. (2022)	All	i-p	No	b-e-nn	c(4)	s
Al-Shabandar et al. (2019)	1 course	i	2 periods	b-e-nn	c(4)	s
Azizah, Pujianto, and Nugraha (2018)		d-i	No	b-ts	c(2)	s
Haiyang et al. (2018)	All	i	Daily	b-nn	c(2)	s
Hassan et al. (2019)	All	i	20 weeks	b-e	c(2)	wdt
Heuer and Breiter (2018)	All	d-i-p	No	b-e	c(2)	s
Hlosta, Zdrahal, and Zendulka (2017)	4 courses	d-i-p	Daily	b-e	c(2)	wdt-c
Hussain et al. (2018)	1 course	d-i-p	No	b-e	c(2)	c
Jha, Ghergulescu, and Moldovan (2019)		d-i-p	No	b-e-nn	r/c(4)	f
Peach et al. (2019)		p	Yes	ts		s
Qiao and Hu (2020)	3 courses	d-i	No	b-e-nn	c(2)	c
Rizvi, Rienties, and Khoja (2019)	4 courses	d-p	No	b	c(3)	c
Tomasevic, Gvozdenovic, and Vranes (2020)	1 course	d-i-p	6 periods	b-nn	c(2)	wdt-c
Waheed et al. (2020)	All	d-i-p	4 quarters	b-e-nn	c(2)	s
This research work	All	i	41 weeks	b-e-nn	c(2)	wdt-c

Note: Scale: Number of courses. Feature: Demographic, interaction, performance. Early prediction/Period. Method: Baseline, ensemble, neuron networks, time series. Technique: Classifier (number of classes), regressor. Model: Single, per week/day/time, per course, per feature.

Our state-of-the-art review revealed that all the student-learning information available in OULAD for the predictive model was never used in some papers. The information was summarised with the three OULAD categories under the column heading "Feature", in Table 1: demographic, interaction and performance. Approximate accuracies of 84%–95% in the last week were reported in the papers in which only student VLE interactions (i) were used (Al-Shabandar et al. 2019; Haiyang et al. 2018; Hassan et al. 2019). Generally speaking, performance improved 5%–8% in those papers that contained supplementary information on the intermediate assessments (p) of the students (Adnan et al. 2022; Peach et al. 2019). No noticeable model performance improvements were evident in the works that used demographic information (d), which only yielded accuracies of around 60% (Jha, Ghergulescu, and Moldovan 2019). In some papers where demographic characteristics were analysed in detail, the highest educational level, region and IMD band were highlighted as the most influential attributes (Rizvi, Rienties, and Khoja 2019). A detailed analysis of "region" highlighted its importance for the detection of dropouts. Some results justify this high relevance, because enrollment on the courses under consideration is only free in certain regions.

In some papers, the early-period component was added to the experiments, to create specific models with time-sorted subsets of the data (column "Early/Period" in Table 1). As expected, the model performance improved as the course progressed. Time series (ts) are one of the specialised methods for dealing with this type of problem (Azizah, Pujianto, and Nugraha 2018; Peach et al. 2019). In other studies with methods based on the construction of classifiers (baseline, ensemble and neural networks), multiple models were generated by grouping the information in temporal windows of days (Haiyang et al. 2018; Hlosta, Zdrahal, and Zendulka 2017), weeks (Hassan et al. 2019) and specific points in time during the course (Al-Shabandar et al. 2019; Tomasevic, Gvozdenovic, and Vranes 2020; Waheed et al. 2020). Rather than considering the time component in the model, all the student interactions were used in many other studies (Adnan et al. 2022; Azizah, Pujianto, and Nugraha 2018; Heuer and Breiter 2018; Hussain et al. 2018; Jha, Ghergulescu, and Moldovan 2019; Qiao and Hu 2020; Rizvi, Rienties, and Khoja 2019) to generate the model.

Numerous methods were considered in the generation of the models (column "Method" in Table 1), so we grouped them into the following categories: baseline, ensemble, neural networks and time series. Ensembling is a technique that combines several models into a larger model, which produces better results than any of its constituents. Ensembling techniques usually take a longer time in ML model training, but even a simple model trained using the ensemble technique can outperform more advanced and specialised models. Qiao and Hu (2020) by comparing the proposed joint neural network model with existing methods of model/feature combination, demonstrated the benefits of applying multiple information sources for prediction (demographic and interactions) and that the proposed joint modelling approach is competent in making collective use of multiple information sources.

The most commonly used supervised machine-learning techniques, under the column heading "Technique" in Table 1, were binary classifiers (c) where the class to be predicted

was either student passes or fails on the course. In other papers, the number of categories was increased to four, looking for excellent students and dropouts (Adnan et al. 2022; Al-Shabandar et al. 2019; Jha, Ghergulescu, and Moldovan 2019). Multi-categorisation, which emerged to counter the imbalance problem, was used to force a balance through pre-processing techniques (e.g., SMOTE (Chawla et al. 2002)). In only one paper (Jha, Ghergulescu, and Moldovan 2019) amongst all those that were reviewed was the problem presented in terms of regressors (r).

The column heading "Model" in Table 1 indicates whether the model design is simple, in the case of models designed with all the data together (Adnan et al. 2022; Al-Shabandar et al. 2019; Azizah, Pujianto, and Nugraha 2018; Haiyang et al. 2018; Heuer and Breiter 2018; Peach et al. 2019), or if multiple models were created grouped by weeks/days/time (wdt) (Hassan et al. 2019; Hlosta, Zdrahal, and Zendulka 2017; Tomasevic, Gvozdenovic, and Vranes 2020), courses (c) Hlosta, Zdrahal, and Zendulka (2017), Rizvi, Rienties, and Khoja (2019), Qiao and Hu (2020), Tomasevic, Gvozdenovic, and Vranes (2020) and/or different characteristics (f) (Jha, Ghergulescu, and Moldovan 2019). Creating multiple models according to some temporal criterion of course duration (i.e., days, weeks, and periods) and analysing them over time helped to address the problem of early detection. Another grouping criterion was the creation of specialised models per course (c) to analyse the variability of their performance in different types of courses or model creation with different feature sets (f), in order to locate the sets that influenced predictive performance more than any others. The combination of modelling criteria created a combinatorial explosion that could only be processed with statistical techniques.

2.3 | Research Gaps

Based on this preliminary study, several lines of work were identified for further research. Analysing significant features for creating predictive models, it was revealed in the study of Adnan et al. (2022) that the course was one of the most relevant features. Therefore, models could be specialised at the course level. In our literature review, no use of this specialisation strategy was shown under the column heading "Model" in many papers, and in those where it was used, work was not done on a large scale with all the available OULAD courses.

Furthermore, the creation of multiple specialised models at the course level plus time-stamped data for early detection specialisation, coupled with large-scale experimentation, caused a combinatorial explosion of models that resulted in excessively high computational costs for their validation.

A remarkable fact is that course data in the domains of both SS and STEM are stored in the OULAD large-scale public data set stores. Although this information was available in the introductory article to the data set (Kuzilek, Hlosta, and Zdrahal 2017), it was not included in any field of its data files. A fact that might perhaps help explain the absence of any studies on the interaction of both types of courses and their relation with student performance over time.

Considering these research gaps, the characteristics of our work are contextualised in the last row of Table 1. The main goal of the experiment was to analyse student logs extracted from the large-scale OULAD data set with multiple supervised classification algorithms, to predict student performance and to compare their accuracies and rankings during academic courses, from the point of view of research into learning analytics, in the particular context of online SS vs. STEM domains. Our three main research questions were therefore as follows:

RQ1. *Which supervised classification algorithm yields the highest prediction performance for predicting student success rates?*

RQ2. *Will the prediction performance of the different supervised classification algorithms differ, depending on the knowledge domain (SS vs. STEM) of the course?*

RQ3. *Will the prediction performance of the different supervised classification algorithms differ, depending on the algorithm, the knowledge domain and the week of the course (early detection)?*

3 | Methods

In this section, the OULAD data set, its settings and the participants in the study are all described. The data analysis procedure is then described, including the pre-processing techniques that are applied, details of the data-mining classifiers in use, and the statistical techniques for the validation of the results. Both the Python and R scripts used, as well as the original data, preprocessed files, and classifier results, are available for experiment replication.[1]

3.1 | OULAD Data Set Structure

The OULAD repository is an Open University (OU) data set created to support learning analytics research (Kuzilek, Hlosta, and Zdrahal 2017). The data set is freely available at https://analyse.kmi.open.ac.uk/open_dataset (CCBY 4.0 licence). The description of CSV data files and their structure and inherent relationships (i.e., primary key, foreign keys, etc.) is thoroughly described.[2] It includes the information on 32,593 students following 22 courses between 2013 and 2014. The standard duration of each course is 9 months and the course data are sorted into three main classes: demographic, learning interaction and performance data. These data are collected in seven files, following an entity-relationship model, although the data are physically available in the form of CSV files.

The following information is sorted in OULAD on the basis of student demographic characteristics: gender, Index of Multiple Deprivation band (IMD band), highest educational level, age, region in which the student lives and disability. This information is shown in the `studentInfo` file.

The VLE interactions of students consisted of 10,655,280 rows with one column, showing the number of times a student interacted with the learning objects. Student interactions (i.e., `studentVle` file) links to the VLE (i.e., `vle` file) using identifier columns showed the following information: the identifier of one

TABLE 2 | Summary of student performance data organised by online courses.

Domain	Module	#Presentation	#Stud.Fail	#Stud.Pass	#Students	%Fail	%Pass
SS	AAA	2	217	531	748	29.01	70.99
SS	BBB	4	4155	3754	7909	52.54	47.46
STEM	CCC	2	2756	1678	4434	62.16	37.84
STEM	DDD	4	3662	2610	6272	58.39	41.61
STEM	EEE	3	1284	1650	2934	43.76	56.24
STEM	FFF	4	4114	3648	7762	53.00	47.00
SS	GGG	3	1020	1514	2534	40.25	59.75
Total SS		9	5392	5799	11,191	48.18	51.82
Total STEM		13	11,816	9586	21,402	55.21	44.79
Total		22	17,208	15,385	32,593	52.80	47.20

of its 22 courses (module presentations), the identifier of one of its 20 types of learning activities, and the time period in which the learning material was available.

The pedagogical design of each online course was independent, such that each course had different types of learning activities and different weighted evaluations, although all the courses had a final exam. The assessment file contains information on assessments during the course (i.e., module and presentation), although the *"final exams are usually missing since they are scored and used for the final marking"* (Kuzilek, Hlosta, and Zdrahal 2017). The assessment results are collected in the `studentsAssessment` file. In our case, we are only interested in that final result, linked to the student and presentation. The final mark reflects the results of the assessments and the final result for the course, yielding a global performance marker, which is included in the `studentInfo` file.

On the basis of the above, the only focus was on student interactions, the main source of the data for predicting the final mark.

3.2 | Setting and Participants

The OULAD data set contains anonymised data sets and clickstream data on learner–VLE interactions, along with course assessment outcomes (Kuzilek, Hlosta, and Zdrahal 2017).

The OULAD data had to undergo specific pre-processing, before the data were ready for our experimental design. The course domain information, either SS or STEM, available in OULAD, was one of the fundamental pillars on which the experimental design was focused. This specific information, although not included in the CSV files, was explicitly associated with each module in Kuzilek, Hlosta, and Zdrahal (2017).

In the design of our large-scale experimentation, we used all the courses and their interaction records (10,655,280) of the students in the different learning tasks. The data were divided into two sets (i.e., training and test), to measure classifier performance, applying the 10-fold cross-validation technique five times. Our experiment was based on binary classification methods with two groups, Pass vs. Fail. The experiment was conducted on a hardware configuration with a 16-core Intel Xeon processor, 128 GB RAM, with 3 Nvidia Titan XP GPUs.

A descriptive summary of the data set, using two classes, Pass and Fail, is shown in Table 2. The Domain column represents the two domain categories covered in the OULAD data set (SS vs. STEM). The #Module column defines the subject of the online course. The #Presentation column represents the number of editions of that online course. #Stud.Fail is the number of student fails on the course and #Stud.Pass is the number of student passes on the course. The remaining #Students, %Fail and %Pass columns correspond to aggregates of #Stud.Pass and #Stud.Fail. Table 2 shows an acceptable balance of both domain and assessment data (52.80% vs. 47.20%).

3.3 | Data Analysis

In this subsection, we detail the pre-processing steps applied to the OULAD data set. First, the data were joined, as we needed the aggregated sum of clicks in the VLE logs of the students enrolled on each course and for each week. Second, we applied different classifiers per course to this transformed data, especially taking into account the problem of early prediction in the first weeks of the course. Finally, we statistically compared the prediction results from different point of views, in response to the research questions.

3.3.1 | Data Preprocessing

In the first step, the `courses.csv` file data were collected, compiling the combination of module code (e.g., AAA, BBC, etc.) and presentation code (e.g., 2013J, 2013B, 2014B, 2014J, etc.). We named this combination of codes as a course, for the sake of simplicity, yielding a total number of 22 courses with the format module_presentation (e.g., AAA_2013J).

Subsequently, all the VLE records were collected from the `studentVle.csv` file, with information related to module, presentation, student, material identifier (learning object), date and sum of clicks. Additional information was extracted from the `vle.csv`

file with the data related to each learning object, mixing the data with each previous record to amplify the information.

The student information was obtained from the `studentinfo.csv` file. Our particular interest was to highlight the final mark in the file, which is the main target of our prediction with four possible values: pass, distinction, fail and withdraw. The final marks of the students were transformed into a binary classification problem, considering *Distinction* as a *Pass* and *Withdrawal* as a *Fail*, transforming the four-class prediction problem into a binary prediction problem, with only two classes: *Pass* vs. *Fail*.

We took the mean value for all courses as a reference (i.e., 38.42 weeks) for the duration of a typical course in weeks, although a 40-week course, divided into 10-week periods, was used in comparisons for homogeneity. We then took the following steps with the pre-processed data, for each course (i.e., as a combination of module and presentation), and for each week in the period from the start (week 0) to the end (week 40) of the course:

- Logs were taken from the VLE with additional student information.
- These logs (number of clicks) were aggregated for each activity type (i.e., learning object type) to the current week.
- A file was generated per course and week with the aggregated click data (e.g., a `data_AAA_2013J_18.csv` file where AAA was the module, 2013J was the presentation and 18 was the week).

As the result of this process, we obtained a total of 902 CSV files (i.e., 22 courses × 41 weeks), as can be seen in Figure 1. It depicts the connections between key fields of the CSV files. Connecting arrows are shown on the left-hand side, linking its relational model. As a result of the OULAD data set pre-processing, each one of the resulting files contained one row *per* student, with the aggregated number of clicks *per* student for each activity type (learning object type), for any given course in any given week.

3.3.2 | Mining Early Performance Prediction

We then selected a well-known Python library, `scikit-learn` (Buitinck et al. 2013; Pedregosa et al. 2011), for the ML phase.

More concretely, we selected 12 classifiers, grouped under the headings: baseline, ensembles and neural networks. We used the name of each classifier class in Scikit-learn (removing the suffix Classifier in some cases, for the sake of brevity) to label the classifiers:

1. Baseline

- DecisionTree: a non-parametric supervised learning method with which a model is created that predicts the value of a target variable by learning simple decision rules inferred from the data features.
- GaussianNB: based on applying Bayes' theorem with the "naive" assumption of conditional independence between each pair of features, given the value of the class variable.
- KNeighbors: computed from a simple majority vote of the nearest neighbours of each point. A query point is assigned a data class label, on the basis of the labels of its nearest neighbours.
- LinearDiscriminantAnalysis (LDA): using a linear decision boundary, generated by fitting class conditional densities to the data and using Bayes' rule, the model fits a Gaussian density to each class, assuming that all classes share the same covariance matrix.
- LogisticRegression: a linear model for classification. Also known as logit regression, maximum-entropy classification, and the log-linear classifier, the probabilities of all

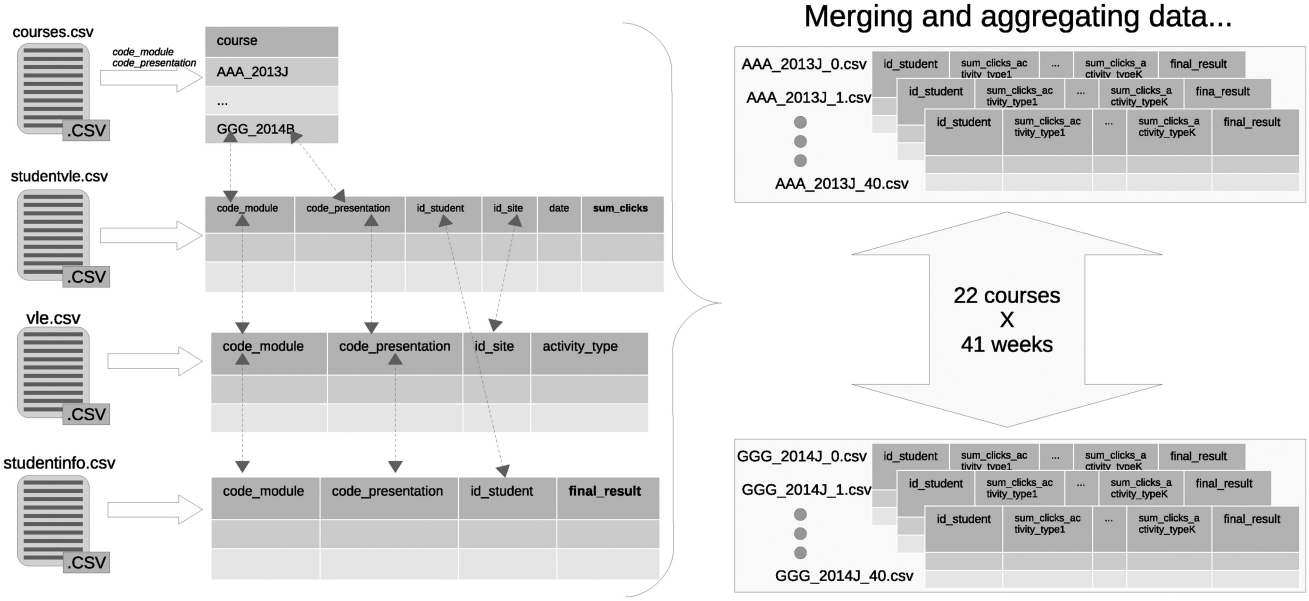

FIGURE 1 | OULAD data set preprocessing obtaining aggregated clicks of students per course and week.

possible outcomes of a single trial are modelled using a logistic function.

- Support Vector Classification (SVC): an implementation of Support Vector Machine (SVM), its training involves optimisation of a convex cost function. The SVM is the best known class of a set of algorithms that use kernel substitution, broadly referred to as kernel methods.

2. Ensembles

- AdaBoost: fits a sequence of weak learners (i.e., models that are only slightly better than random guessing, such as small decision trees) on repeatedly modified versions of the data. The predictions from all of them are then combined through a weighted majority vote (or sum) to produce the final prediction.
- Bagging: forms a class of algorithms which build several instances of a black-box estimator on random subsets of the original training set and then their individual predictions are aggregated to form a final prediction. These methods are used as a way of reducing the variance of a base estimator (e.g., a decision tree), by introducing randomisation into its construction procedure and then making an ensemble out of it.
- ExtraTrees: implements a meta estimator that fits a number of randomised decision trees (a.k.a. extra-trees) on various sub-samples of a data set, after which averaging is applied to improve the predictive accuracy and control over-fitting.
- RandomForest: each tree in the ensemble is built from a sample drawn with replacement (i.e., a bootstrap sample) from the training set. When splitting each node during the construction of a tree, the best split is found either from all input features or a random subset.
- XGBoost: implementation of parallel tree boosting (also known as GBDT, GBM) that solves many data science problems in a fast and accurate way.

3. Neural networks

- Multi-layer Perceptron (MLP): a supervised learning algorithm that learns a function by training on a data set, where m is the number of dimensions for input and o is the number of dimensions for output. Given a set of features and a target, an MLP can learn a non-linear function approximator for either classification or regression. It differs from logistic regression, in so far as there can be one or more non-linear layers, called hidden layers, between the input and the output layer.

We encoded each categorical value as a numerical value of zero or one, for each prediction value. In all cases, we trained the classifiers with default hyper-parameter values and validated each one for each combination of course and week. To do so, we applied a pipeline with the following two steps: *SMOTE*, to avoid the problem of unbalanced data sets, and *StandardNormalization*, to optimise the performance of the classifiers working with normalised numeric data.

A 10-fold cross-validation training was repeated five times, using the above pipeline, that yielded the average level of accuracy. We therefore applied our 902 training sets (i.e., course records per week for each learner on learning objects) and the 12-classifier training that yielded a total of 10,824 models with their corresponding mean accuracy. The complete process is shown in Figure 2.

Finally, for each row obtained, according to the type of the course, we added the corresponding domain Social Sciences

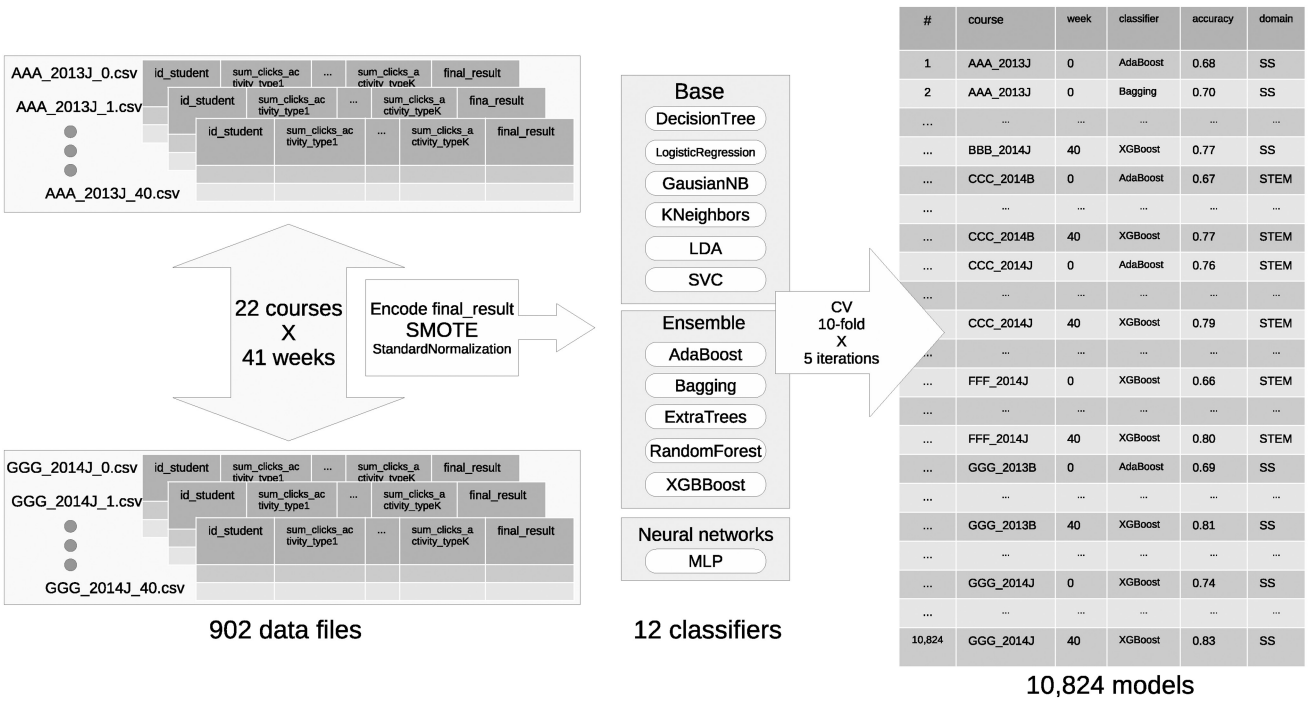

FIGURE 2 | Training process obtaining mean accuracy per model using the new domain variable (SS vs. STEM).

(SS) for the courses in modules AAA, BBB and GGG, and Science, Technology, Engineering and Mathematics (STEM) for the rest of the courses. Although the domain information was not directly available in the OULAD CSV files for its extraction, the data set description (Kuzilek, Hlosta, and Zdrahal 2017) established the matching between modules and domain. With all this information, we were then ready to address the research questions.

3.3.3 | Statistical Comparisons

The prediction performance of student success rates (i.e., model accuracy) was compared to respond to the three research questions (see Section 2.3), using the following approach: statistical difference in accuracy between different supervised classification algorithms (RQ1); differences in model accuracy associated with the knowledge domain (SS vs. STEM) (RQ2); and differences associated with early detection (week of course), domain and supervised classification algorithms (RQ3) using beta regression through the *betareg* package (Cribari-Neto and Zeileis 2010). Statistical differences for rank position between supervised classification algorithm performance and week of course were tested using Friedman's Two-Way Analysis of Variance by Ranks to compare matched samples, with a significance level of <0.001 (RQ3). Beta-regression modelling was required for modelling beta-distributed data (i.e., intervals between 0 and 1) such as rates and proportions. Post hoc analysis (tukey) was calculated based on the contrasts for all pairwise comparisons amongst the estimated marginal means, with the *emmeans* package (Lenth 2022). Visual comparisons of model performance were drawn using the *ggplot2* package (Wickham 2016), under the R programming language (R Core Team 2022). We used SPSS Release 28.0 (IBM Corp. 2021), for running the Friedman test.

4 | Results

In this section, each of the different research questions are addressed. Following the methodology set out in the previous section, the following results were achieved.

RQ1. *Which supervised classification algorithm yields the highest prediction performance for predicting student success rates?*

In general, the accuracy of the classifiers (p-value <0.001) varied significantly. In Figure 3, a gradient in algorithm performance was evident. Post hoc analysis, based on contrasts for all pairwise comparisons, mainly showed three groups of classifiers. The first group (those with the letter *a* in the figure) formed of the classifiers *Random Forest* and *Extra Trees* showed the highest predictive performance. The second group formed of five classifiers (those with the letter *a,b* and *b,c* in Figure 3) showed an intermediate level of predictive performance (although still with high accuracy values). Finally, the last group of classifiers (those with the letter *c* in the Figure 3) represented the worst predictive performance over all the classifiers. Interestingly, the *Gaussian Naive Bayes* and *Decision Tree* classifiers yielded the lowest scores (Figure 3).

Following the results of Friedman's Analysis, we ordered the classifiers according to their ranking in the test (the highest values are the highest ranked), where the top five classifiers were *RandomForest* (ensemble), *ExtraTrees* (ensemble), *SVC* (baseline), *MLP* (neural networks) and *AdaBoost* (ensemble), as can be seen in Table 3.

RQ2. *Will the prediction performance of the different supervised classification algorithms differ, depending on the knowledge domain (SS vs. STEM) of the course?*

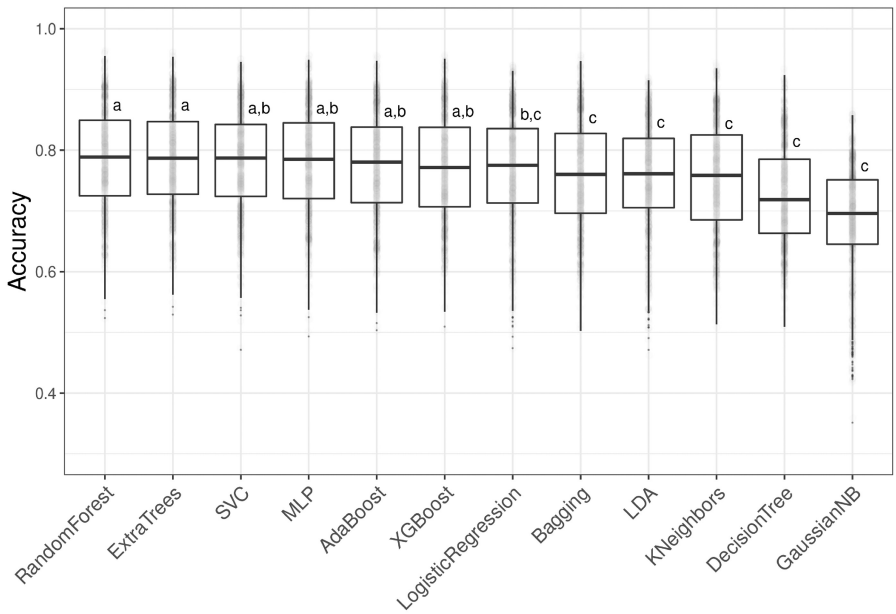

FIGURE 3 | Differences in accuracy between different classifiers. Post hoc pair-wise comparisons are highlighted as letters above each boxplot. Boxplots connected with different letters were statistically different.

The analysis of accuracy variability over the two domains showed that the STEM domain had an average accuracy value of (0.79 ± 0.09), which was significantly higher (p value < 0.001) than the average accuracy for the SS domain (0.74 ± 0.10). Besides, the range of variability was slightly lower (although not significant) in the STEM domain than in the SS domain. Thus, classifier performance yielded higher accuracy levels and lower variability for STEM courses than for SS courses (Figure 4).

Differences in accuracy values associated with the domain across different classifiers revealed that these differences were significant (p value < 0.001) for all the classifiers, although not all the classifiers showed the same range of differences. Notably, the *Gaussian Naive Bayes* classifier showed the highest difference in terms of accuracy with respect to course domain (SS vs. STEM; Figure 5).

TABLE 3 | Mean rank per classifier.

Classifier	Mean rank
RandomForest	**10.04**
ExtraTrees	**9.59**
SVC	**9.35**
MLP	**9.35**
AdaBoost	**7.99**
LogisticRegression	7.30
XGBoost	6.57
LinearDiscriminantAnalysis	5.03
Bagging	4.70
KNeighbors	4.12
DecisionTree	2.27
GaussianNB	1.70

Note: Top 5 values in bold.

RQ3. *Will the prediction performance of the different supervised classification algorithms differ, depending on the algorithm, the knowledge domain and the week of the course (early detection)?*

In general, prediction performance in terms of accuracy increased with each week of the course (higher accuracy for last weeks of the course) for both course domains (SS vs. STEM); however, this trend was less marked in SS than in STEM (Figure 6). Significant differences associated with course domain were consistent across the five periods of time (i.e., weeks) (p value < 0.001). Differences between SS and STEM were higher at the start of the course (week 0 and 10) and tended to decrease at the last week (week 40), even though, they were always significant (interaction between domain and week was not significant, p value > 0.05). Differences associated with the classifiers were also significant as the weeks of the course advanced for both domains (p value < 0.05) where *Gaussian Naive Bayes* and *Decision Tree* had the lowest prediction performance of all the classifiers. Despite the variability in accuracy associated with each classifier, it is interesting to mention that some points were out of range (marked as dots) and a contrasting distribution of these points was evident. For example, the high spread of accuracy variability in STEM, when present, only appeared in the positive (high values) of the accuracy variability. In contrast, in SS courses many more points were presented, always within the lower part of the accuracy variability (Figure 6).

We used the mean rank for 10-week periods, to order the classifiers from another point of view, thereby evaluating classifier performance over the study period, as can be seen in Table 4, highlighting the top 5 values. According to this ranking, the top five classifiers were *RandomForest* (ensemble), *ExtraTrees*

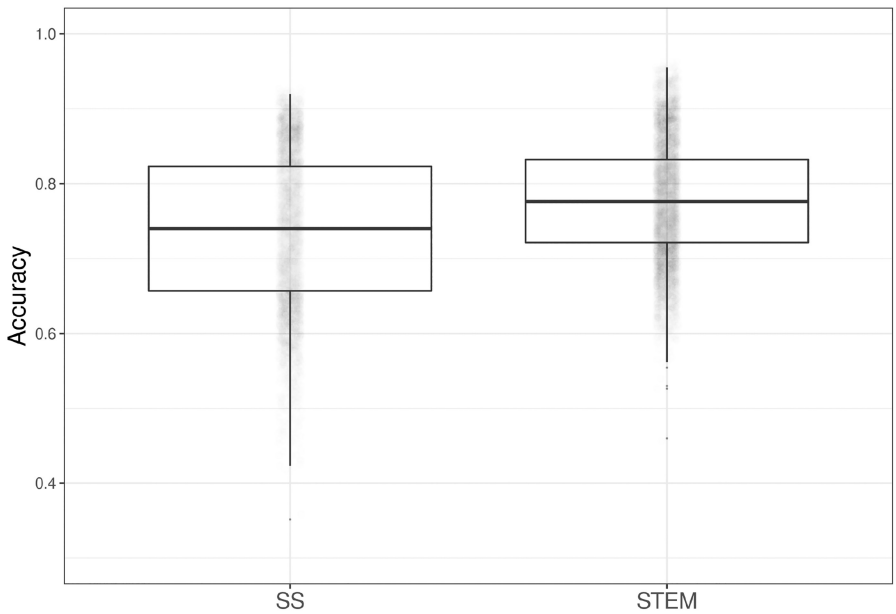

FIGURE 4 | Differences in accuracy (all classifiers) for each domain (SS vs. STEM). Boxplots represent the median and the interquartile range. Differences were significant ($p < 0.001$) at $\alpha = 0.05$.

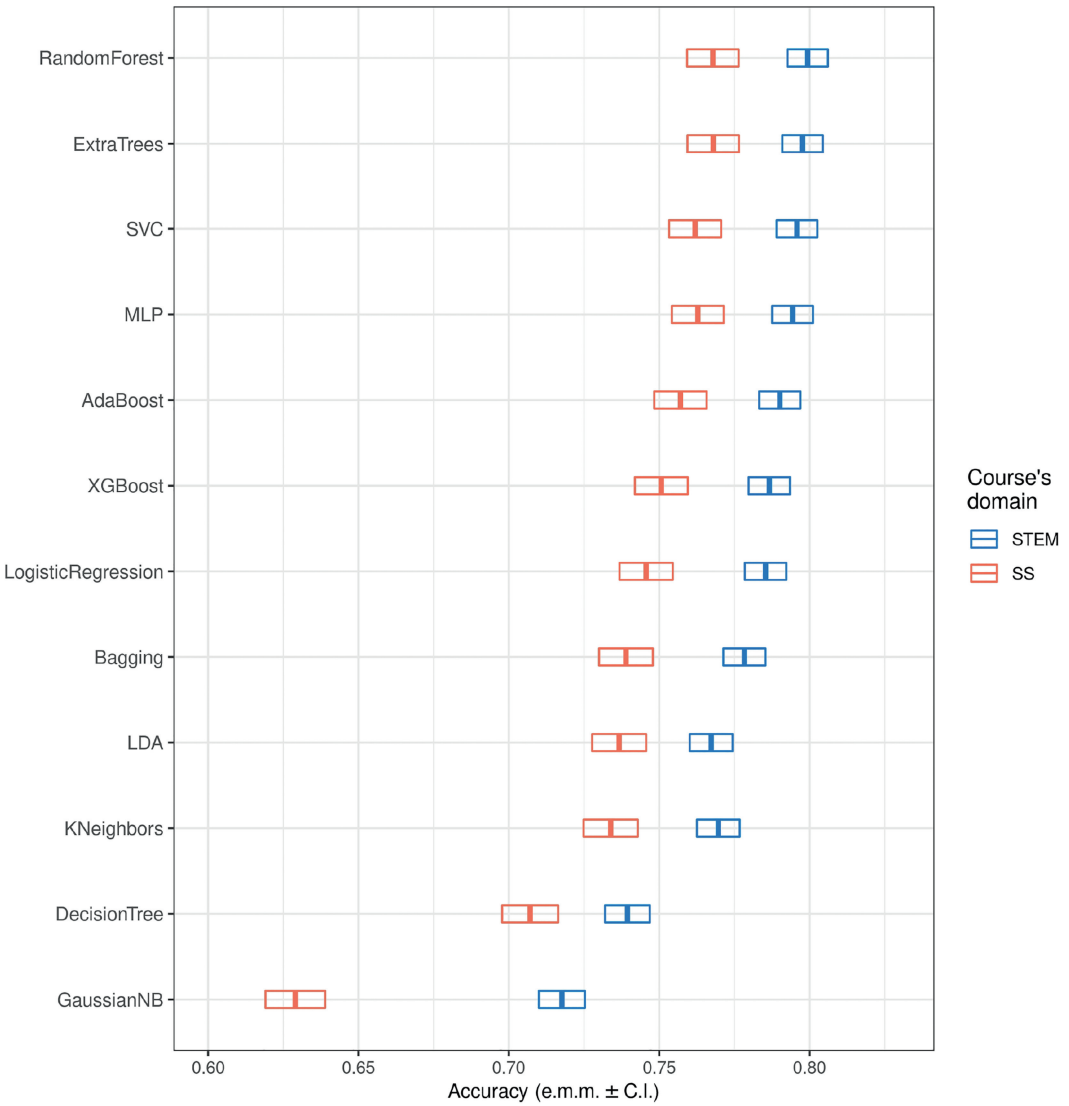

FIGURE 5 | Average accuracy variability for different classification algorithms with respect to course domain (SS vs. STEM). C.I. = confidence intervals; e.m.m. = estimated marginal means.

(ensemble), *SVC* (baseline), *MLP* (neural networks) and *AdaBoost* (ensemble). Table 5 shows the differences between the classifiers across all weeks. Significant results using Friedman's test led us to reject the null hypothesis of equal performance of classifiers for all time periods.

5 | Discussion

Studies are scarce on the impact of the domain (SS vs. STEM) when predicting student outcomes using data on large-scale online and automated courses. Our results provide evidence that the type of course knowledge domain (SS vs. STEM) influences the prediction of learning outcomes, as can be appreciated in our results, with lower variability for STEM. While differences have been suggested in the literature on online learning by Finnegan, Morris, and Lee (2008), our study is the first to advance an empirical test of the logged interactions of 32,593 students following 22 (9 SS and 13 STEM) online courses stored on the OU learning analytics database.

These results are relevant for decision-making in the management of online courses. In particular, the domain of the courses under study is a factor that is worth considering when building models for drop-out risk and performance prediction.

Bearing in mind the available data, the main baseline in our design of performance prediction models was to only use data from data records of interaction with learning objects (i.e., logs). The accuracy of this minimalist modelling strategy has been experimentally demonstrated in the detection of student performance failure (Riestra-González, del Puerto Paule-Ruíz, and Ortin 2021). Another advantage of this minimalist approach was that the results and models could be extrapolated to any online course design regardless of the platform where it was implemented. This advantage was applied by Al-Shabandar et al. (2019) in their experimental designs when they built a single prediction model that fitted the learning interaction of two large-scale data sets such as OULAD (Kuzilek, Hlosta, and Zdrahal 2017) and HarvardX/MITx (Ho et al. 2014).

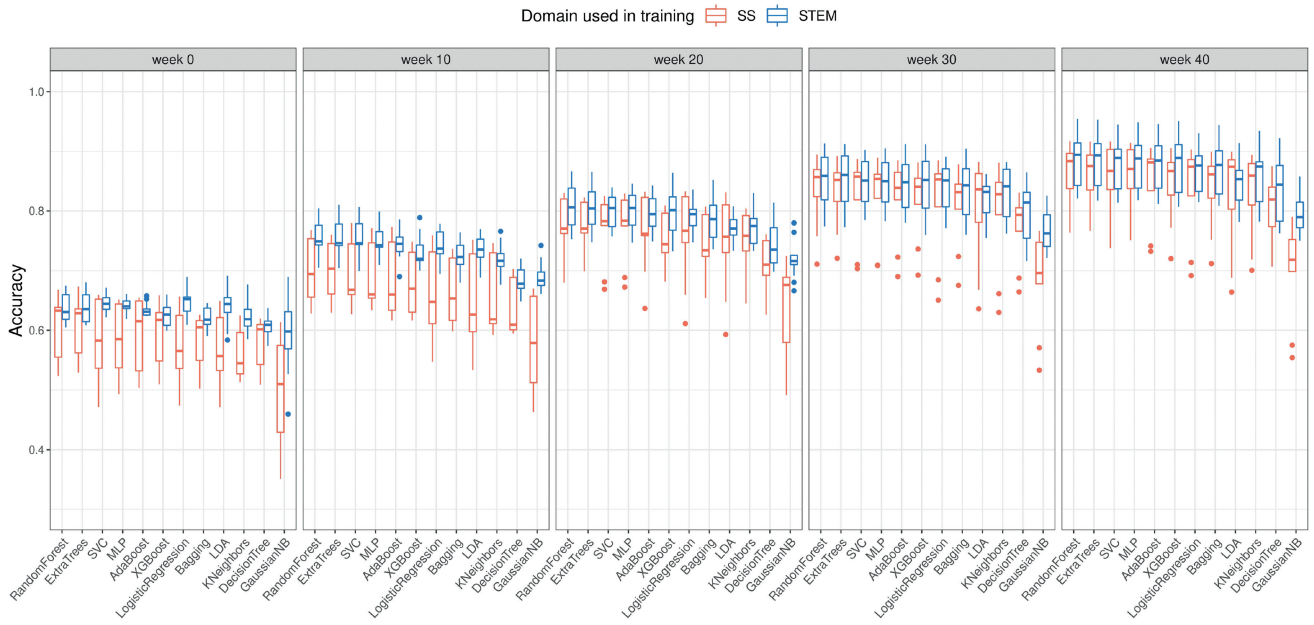

FIGURE 6 | Accuracy variability amongst classifiers for each domain (SS vs. STEM) and across different weeks of the course.

TABLE 4 | Rank for classifiers per week (in alphabetical order).

Classifier/Week	0	10	20	30	40
AdaBoost	7.68	**8.18**	**8.14**	**7.59**	**7.82**
Bagging	4.64	4.09	4.86	5.14	5.09
DecisionTree	3.18	1.82	1.95	2.14	2.23
ExtraTrees	**8.68**	**10.18**	**9.55**	**10.36**	**10.00**
GaussianNB	3.05	1.73	1.27	1.23	1.18
KNeighbors	4.36	3.91	4.05	4.09	4.23
LDA	6.82	5.05	4.68	3.95	3.95
LogisticRegression	**7.82**	7.09	7.55	6.82	6.50
MLP	**8.59**	**9.36**	**9.73**	**9.05**	**9.36**
RandomForest	**8.00**	**10.14**	**10.36**	**11.00**	**11.36**
SVC	**8.91**	**10.45**	**9.18**	**9.14**	**9.05**
XGBoost	6.27	6.00	6.68	7.50	7.23

Note: Bold marks the 5th best performance per week.

TABLE 5 | Friedman's test per week.

Week	*N*	Chi-square	df	Asymp.Sig
0	22	87.077	11	<0.001
10	22	191.867	11	<0.001
20	22	178.720	11	<0.001
30	22	189.343	11	<0.001
40	22	191.266	11	<0.001

From an instructor's perspective, the ability to predict student performance using interaction logs is a powerful tool in modern education. Here, we have proposed an analytical design that is easily implemented and easily compared for assessing the detection of student performance failure. With our experimental accuracy results, we recommend that instructors identify the low performance in week 10 of 40 in the STEM courses and from week 20 of 40 in the SS courses. The low early performance of students suggests providing timely feedback and fostering engagement. Instructors should ensure course accessibility with adaptive tools, adjusting feedback based on the individual student's progress, needs or abilities. On the other hand, academic advisors should identify early at-risk students by collaborating with instructors, offering personalised guidance on study strategies and time management and connecting students to mentoring programmes and workshops. They can help build support networks through peer mentoring and study groups while collecting student feedback to inform institutional improvements, creating a supportive and effective learning environment.

Two other important design decisions, as well as only constructing log-based models, were drawn from the results section of the related work, Section 2.3. The first was to build course-specific models and the second was to work on a large scale with many courses and many students enrolled on each course. Models of higher accuracy when the predictions were separately customised for each course were shown in the work of Wolff et al. (2014). In some previous papers, experimentation with data from online and blended courses was described, conducted in small-size universities (Manzanares et al. 2017). Data extraction with open-source software, UBUMonitor (Marticorena-Sánchez et al. 2022) in this case, compiled student interactions and grading records from Moodle, a well-known VLE. These results from small-scale empirical experimentation were very limited to their specific experimental conditions, which in many cases, prevented their validation under different conditions and course designs, as neither the pedagogical design nor the socioeconomic context of the students following the course were known. Having a large, reliable open data set, such as OULAD, anonymised and legally

compliant to design and experiment with prediction models, is a requirement to create an experimental framework that multiple researchers can share. Previous work on the reference data set helped us to corroborate the experimental results and to manage the evolution of prediction improvements with the design and the evaluation of new experiments (Adnan et al. 2022). The models may be adjusted in specific ways following the extrapolation of the generic results obtained in this work and their application to specific, smaller-sized contexts.

Experimentation in the area of EDM and LA needs to be done through experimental work. Thirteen ML algorithms were employed to build a total of 10,824 prediction models. The algorithms included the three main categories of Basic, Ensemble and Neural Networks. The results of comparing the predictive accuracy of student success or failure models showed that (i) RandomForest, ExtraTrees, SVC, MLP, AdaBoost and XGBoost yielded the highest accuracy. A ranking and a grouping was established to reuse them (see Figure 3 and Table 3); (ii) regardless of the algorithm used, the prediction accuracies of the STEM courses models were always higher (see Figure 5).

We were able to recognise differences between the two branches of knowledge (SS vs. STEM), following an experimental design based on nothing other than logs. Furthermore, the analysis was performed from the start of the course, so that "early detection" could be performed. Some differences also showed that the STEM courses had higher accuracy values and lower variability (i.e., a narrow spread of accuracy values) than the SS courses throughout the weeks of the course, but mostly during the first weeks. It facilitated an early dropout detection policy in this domain, based on nothing other than VLE activity alone. However, over time, the difference was observed to narrow, without becoming equal between the two domains.

Given the high dropout rate in STEM established by Bernacki, Chavez, and Uesbeck (2020) and the differences detected between SS vs. STEM in this work, it can be established that while the use of early prediction based on logs was applicable with greater success in STEM, in the case of SS, it should be accompanied by more exhaustive analysis. Some avenues to do so might be to add other social data or in-depth personalised follow-up of the student cohort through interviews or surveys at the start of the course.

From the Table 1 with the comparison of related works in Section 2, we provide the concrete benchmark accuracy that serves to compare and evaluate this type of student early dropout prediction models, along with the assumptions made in their construction. Of the 10,824 models built, our best accuracy was 0.9547, in the last week of the course. Al-Shabandar et al. (2019), using only log data from the last week of a course, report a maximum accuracy of 0.95, which is equivalent to our result. Riestra-González, del Puerto Paule-Ruíz, and Ortin (2021), using only log data, report their best accuracy of 0.90 with data obtained in the middle of the course. Tomasevic, Gvozdenovic, and Vranes (2020), in addition to interaction data, use student demographic and grade data and report a best performance F-measure of 96.6%. Adnan et al. (2022), constructing four-category classifiers and pooling interaction and grade data from all courses, report their best accuracy of 0.83 in the last week of the course.

Regarding the explainability of model features, we have previously justified and discussed our assumption to consider only interaction features of the students in the construction of the prediction models without taking into account demographics and intermediate grades. On this premise, we would advise against trying to find a generalist explanation of the most important interaction features that the model uses to decide, due to several issues. First, the disparity of features between courses would reduce the scope of such results, as they would not be generalisable. Second, the scarce semantic documentation on the features (Kuzilek, Hlosta, and Zdrahal 2017) that make up the courses would limit our ability to reason about the possible results obtained. This does not prevent us from suggesting this line of future work, but through courses, with a better understanding of the pedagogical design of the activities and resources used.

6 | Conclusions

The overall conclusion drawn from the results of this experimental study has been that student success can be predicted with high accuracy using VLE log information through ML techniques. This conclusion is concordant with the results of the literature reviewed in this study. Experimentation with the large-scale data set, along with analysis of 13 experimental research papers on OULAD, concluded that model accuracy slightly improved after adding intermediate grading features and demographic information. Extracting student access logs to the VLE provided sufficient data with which to build very accurate models. Moreover, these models could be generalisable to different VLEs and courses. The alternative of constructing models with more student features achieved greater accuracy, but was less generalisable. This trade-off analysis between accuracy and generalisability can help stakeholders in their decision-making.

We advanced the design criteria of a framework for building more accurate models by (i) adding a temporal component and (ii) considering the difference in knowledge domains of the courses (SS vs. STEM). In this regard, several contributions and considerations may be advanced. The first was that even if large-scale data were available, models should include information (i.e., logs, grades and socio-economic) related to course and domain (SS vs. STEM). The second is that when building models, the subsets of SS and STEM courses behaved differently in terms of their accuracy and the week of the course influenced prediction. Furthermore, high early-prediction accuracy on SS courses were not noted until almost halfway through the courses, while high levels of accuracy were achieved for STEM within the first few weeks of the course.

Although we found differences between student interaction logs on SS and on STEM courses, the causes were not analysed in depth. The inclusion of models with different levels of interpretation, particularly white-box models, may help stakeholders to interpret the results in both domains and take concrete actions to improve future outcomes. The main line of future work is the search for the causes of the differences between SS vs. STEM courses. The reason why VLE interaction differs between SS and STEM students must be analysed, taking into account the type of learning activities on the course.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are openly available in OULAD-Comparative analysis-SS vs STEM courses at https://data.mendeley.com/datasets/8624n5symc/1, reference number 10.17632/8624n5symc.1.

Endnotes

[1]See https://data.mendeley.com/datasets/8624n5symc.

[2]Its entity-relationship model is available at https://analyse.kmi.open.ac.uk/open_dataset#description.

References

Adnan, M., A. A. S. Alarood, M. I. Uddin, and I. Ur Rehman. 2022. "Utilizing Grid Search Cross-Validation With Adaptive Boosting for Augmenting Performance of Machine Learning Models." *PeerJ Computer Science* 8: e803. https://doi.org/10.7717/peerj-cs.803.

Aldowah, H., H. Al-Samarraie, and W. Fauzy. 2019. "Educational Data Mining and Learning Analytics for 21st Century Higher Education: A Review and Synthesis." *Telematics and Informatics* 37: 13–49. https://doi.org/10.1016/j.tele.2019.01.007.

Alsariera, Y. A., Y. Baashar, G. Alkawsi, A. Mustafa, A. A. Alkahtani, and N. Ali. 2022. "Assessment and Evaluation of Different Machine Learning Algorithms for Predicting Student Performance." *Computational Intelligence and Neuroscience* 2022: 4151487. https://doi.org/10.1155/2022/4151487.

Al-Shabandar, R., A. J. Hussain, P. Liatsis, and R. Keight. 2019. "Detecting At-Risk Students With Early Interventions Using Machine Learning Techniques." *IEEE Access* 7: 149464–149478. https://doi.org/10.1109/ACCESS.2019.2943351.

Azizah, E. N., U. Pujianto, and E. Nugraha. 2018. "Comparative Performance Between C4.5 and Naive Bayes Classifiers in Predicting Student Academic Performance in a Virtual Learning Environment." In *2018 4th International Conference on Education and Technology (ICET)*, edited by D. E. Kusumaningrum, 18–22. Malang, Indonesia: Institute of Electrical and Electronics Engineers Inc.

Bernacki, M. L., M. M. Chavez, and P. M. Uesbeck. 2020. "Predicting Achievement and Providing Support Before STEM Majors Begin to Fail." *Computers & Education* 158: 103999. https://doi.org/10.1016/j.compedu.2020.103999.

Buitinck, L., G. Louppe, M. Blondel, et al. 2013. "API Design for Machine Learning Software: Experiences From the Scikit-Learn Project." In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 108–122. Prague: Springer.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. "SMOTE: Synthetic Minority Over-Sampling Technique." *Journal of Artificial Intelligence Research* 16: 321–357. https://doi.org/10.1613/jair.953.

Cribari-Neto, F., and A. Zeileis. 2010. "Beta Regression in R." *Journal of Statistical Software* 34, no. 2: 1–24. https://doi.org/10.18637/jss.v034.i02.

Finnegan, C., L. V. Morris, and K. Lee. 2008. "Differences by Course Discipline on Student Behavior, Persistence, and Achievement in Online Courses of Undergraduate General Education." *Journal of College Student Retention: Research, Theory & Practice* 10: 39–54. https://doi.org/10.2190/CS.10.1.d.

Gray, C. C., and D. Perkins. 2019. "Utilizing Early Engagement and Machine Learning to Predict Student Outcomes." *Computers & Education* 131: 22–32. https://doi.org/10.1016/j.compedu.2018.12.006.

Haiyang, L., Z. Wang, P. Benachour, and P. Tubman. 2018. "A Time Series Classification Method for Behaviour-Based Dropout Prediction." In *2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT)*, edited by, M. Chang, N.-S. Chen, R. Huang, Kinshuk, K. Moudgalya, S. Murthy, and D. G. Sampson, 191–195. Bombay, India: Institute of Electrical and Electronics Engineers Inc.

Hassan, S., H. Waheed, N. R. Aljohani, M. Ali, S. Ventura, and F. Herrera. 2019. "Virtual Learning Environment to Predict Withdrawal by Leveraging Deep Learning." *International Journal of Intelligent Systems* 34: 1935–1952. https://doi.org/10.1002/int.22129.

Helal, S., J. Li, L. Liu, et al. 2018. "Predicting Academic Performance by Considering Student Heterogeneity." *Knowledge-Based Systems* 161: 134–146. https://doi.org/10.1016/j.knosys.2018.07.042.

Heuer, H., and A. Breiter. 2018. "Student Success Prediction and the Trade-Off Between Big Data and Data Minimization - Digitale Bibliothek - Gesellschaft für Informatik e.V." https://dl.gi.de/handle/20.500.12116/21041.

Hlosta, M., Z. Zdrahal, and J. Zendulka. 2017. "Ouroboros: Early Identification of At-Risk Students Without Models Based on Legacy Data." In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, 6–15. New York, NY: Association for Computing Machinery.

Ho, A. D., J. Reich, S. O. Nesterko, et al. 2014. "HarvardX and MITx: The First Year of Open Online Courses, Fall 2012-Summer 2013." SSRN Electronic Journal. https://doi.org/10.2139/ssrn.2381263.

Hussain, M., W. Zhu, W. Zhang, and S. M. R. Abidi. 2018. "Student Engagement Predictions in an e-Learning System and Their Impact on Student Course Assessment Scores." *Computational Intelligence and Neuroscience* 2018, no. 1: 6347186.

IBM Corp. 2021. "IBM SPSS Statistics for Windows." https://www.ibm.com/analytics/spss-statistics-software/.

Jha, N. I., I. Ghergulescu, and A. N. Moldovan. 2019. "Oulad Mooc Dropout and Result Prediction Using Ensemble, Deep Learning and Regression Techniques." In *CSEDU 2019 - Proceedings of the 11th International Conference on Computer Supported Education*, edited by H. Lane, S. Zvacek, and J. Uhomoibhi, vol. 2, 154–164. Heraklion, Crete, Greece: SciTePress.

Karagiannis, I., and M. Satratzemi. 2018. "An Adaptive Mechanism for Moodle Based on Automatic Detection of Learning Styles." *Education and Information Technologies* 23, no. 3: 1331–1357. https://doi.org/10.1007/s10639-017-9663-5.

Knight, S., A. F. Wise, and B. Chen. 2017. "Time for Change: Why Learning Analytics Needs Temporal Analysis." *Journal of Learning Analytics* 4: 7–17. https://doi.org/10.18608/jla.2017.43.2.

Kuzilek, J., M. Hlosta, and Z. Zdrahal. 2017. "Open University Learning Analytics Dataset." *Scientific Data* 4: 1–8. https://doi.org/10.1038/sdata.2017.171.

Lenth, R. V. 2022. "Emmeans: Estimated Marginal Means, Aka Least-Squares Means." *R Package Version* 1, no. 7: 5. https://CRAN.R-project.org/package=emmeans.

Manzanares, M. C. S., R. M. Sánchez, C. I. G. Osorio, and J. F. Dìez-Pastor. 2017. "How Do B-Learning and Learning Patterns Influence Learning Outcomes?" *Frontiers in Psychology* 8, no. 745: 1–13. https://doi.org/10.3389/fpsyg.2017.00745.

Marticorena-Sánchez, R., C. López-Nozal, Y. P. Ji, C. Pardo-Aguilar, and Á. Arnaiz-González. 2022. "UBUMonitor: An Open-Source Desktop Application for Visual E-Learning Analysis With Moodle." *Electronics* 11, no. 6: 954. https://doi.org/10.3390/electronics11060954.

Peach, R. L., S. N. Yaliraki, D. Lefevre, and M. Barahona. 2019. "Data-Driven Unsupervised Clustering of Online Learner Behaviour." *NPJ Science of Learning* 4: 1–11. https://doi.org/10.1038/s41539-019-0054-0.

Pedregosa, F., G. Varoquaux, A. Gramfort, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12, no. 85: 2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html.

Peña-Ayala, A. 2014. "Educational Data Mining: A Survey and a Data Mining-Based Analysis of Recent Works." *Expert Systems With Applications* 41, no. 4: 1432–1462. https://doi.org/10.1016/j.eswa.2013.08.042.

Qiao, C., and X. Hu. 2020. "A Joint Neural Network Model for Combining Heterogeneous User Data Sources: An Example of At-Risk Student Prediction." *Journal of the Association for Information Science and Technology* 71, no. 10: 1192–1204. https://doi.org/10.1002/asi.24322.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rastrollo-Guerrero, J., J. A. Gomez-Pulido, and A. Domínguez. 2020. "Analyzing and Predicting Students' Performance by Means of Machine Learning: A Review." *Applied Sciences* 10, no. 3: 1042. https://doi.org/10.3390/app10031042.

Riestra-González, M., M. del Puerto Paule-Ruíz, and F. Ortin. 2021. "Massive LMS Log Data Analysis for the Early Prediction of Course-Agnostic Student Performance." *Computers & Education* 163: 104108. https://doi.org/10.1016/j.compedu.2020.104108.

Rivas, A., A. González-Briones, G. Hernández, J. Prieto, and P. Chamoso. 2021. "Artificial Neural Network Analysis of the Academic Performance of Students in Virtual Learning Environments." *Neurocomputing* 423: 713–720. https://doi.org/10.1016/j.neucom.2020.02.125.

Rizvi, S., B. Rienties, and S. A. Khoja. 2019. "The Role of Demographics in Online Learning; A Decision Tree Based Approach." *Computers & Education* 137: 32–47. https://doi.org/10.1016/j.compedu.2019.04.001.

Roll, I., D. M. Russell, and D. Gaševic. 2018. "Learning at Scale." *International Journal of Artificial Intelligence in Education* 28: 471–477. https://doi.org/10.1007/s40593-018-0170-7.

Romero, C., and S. Ventura. 2020. "Educational Data Mining and Learning Analytics: An Updated Survey." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10: e1355. https://doi.org/10.1002/widm.1355.

Romero, C., S. Ventura, and E. García. 2008. "Data Mining in Course Management Systems: Moodle Case Study and Tutorial." *Computers & Education* 51, no. 1: 368–384. https://doi.org/10.1016/j.compedu.2007.05.016.

Roy, K. S., K. Roopkanth, V. U. Teja, V. Bhavana, and J. Priyanka. 2018. "Student Career Prediction Using Advanced Machine Learning Techniques." *International Journal of Engineering and Technology* 7, no. 2.20: 26–29.

Sáiz-Manzanares, M. C., R. Marticorena-Sánchez, N. Muñoz-Rujas, et al. 2021. "Teaching and Learning Styles on Moodle: An Analysis of the Effectiveness of Using STEM and Non-STEM Qualifications From a Gender Perspective." *Sustainability* 13, no. 3: 1166. https://doi.org/10.3390/su13031166.

Tomasevic, N., N. Gvozdenovic, and S. Vranes. 2020. "An Overview and Comparison of Supervised Data Mining Techniques for Student Exam Performance Prediction." *Computers & Education* 143: 103676. https://doi.org/10.1016/j.compedu.2019.103676.

Waheed, H., S. U. Hassan, N. R. Aljohani, J. Hardman, S. Alelyani, and R. Nawaz. 2020. "Predicting Academic Performance of Students From VLE Big Data Using Deep Learning Models." *Computers in Human Behavior* 104: 106189. https://doi.org/10.1016/j.chb.2019.106189.

Wang, M.-T., and J. L. Degol. 2017. "Gender Gap in Science, Technology, Engineering, and Mathematics (STEM): Current Knowledge, Implications for Practice, Policy, and Future Directions." *Educational Psychology Review* 29: 119–140. https://doi.org/10.1007/s10648-015-9355-x.

Wickham, H. 2016. *ggplot2: Elegant Graphics for Data Analysis*. New York, NY: Springer-Verlag. https://ggplot2.tidyverse.org.

Wolff, A., Z. Zdrahal, D. Herrmannova, J. Kuzilek, and M. Hlosta. 2014. "Developing Predictive Models for Early Detection of At-Risk Students on Distance Learning Modules." In *Machine Learning and Learning Analytics Workshop at the 4th International Conference on Learning Analytics and Knowledge (LAK14)*, vol. 1137, 24–28. New York, NY: Association for Computing Machinery.