university of groningen

faculty of science and engineering

Predict Student Achievement Using Virtual Learning Environment Log Data

Eugen-Florian Falca

university of groningen

faculty of science and engineering

University of Groningen

Predict Student Achievement Using Virtual Learning Environment Log Data

Bachelor's Project Thesis

To fulfill the requirements for the Bachelor's Project in Computing Science
at the University of Groningen under the supervision of
Prof. Dr. Asadollah Shahbahrami (Computing Science, University of Groningen)
and
Prof. Dr. Carlo Galuzzi (Computer Engineering, Delft University of Technology)

Eugen-Florian Falca (s4341392)

July 31, 2023

Contents

Abstract

The usage of virtual learning environments (VLE) in tertiary education gave new opportunities to researchers from the Educational Data Mining field. Predictive analytics is a sub-group that can solve problems that affect the students, such as extended duration to finish a degree. Using 3 machine learning algorithms, we analyse which are the aspects that impact student achievement. On the other hand, we predict if a student will pass or fail a course, from the midpoint of the module. The Open University Learning Analytics dataset was used, creating a set of features that applies to all the courses from this data collection. The independent variables are categorised into 3 groups, these being demographics, assessments and VLE interactions. In both experiments, the best results were obtained using Random Forest (RF) and Light Gradient-Boosting Machine. We conclude that when predicting future achievement, the most impactful component is the past results of the student. Furthermore, we obtained an 88.88 Area Under the Curve, using RF, when predicting the student performance, while using data from the first half of the course.

1 Introduction

Education is an essential field since it produces the workforce, that will develop the economy and improve the society of a country. Each year, millions of students join universities around the globe, so that at the end of a long process, to receive a certification that approves their knowledge and expertise within a field. To do this, the student needs to show a certain level of qualification. This measure is determined by the adequate committees and boards within the university, in relation to the ones proposed by the officials of the corresponding country. This bar is used to assess the performance of the student within the academic environment and it determines the student's achievement.

Numerous factors influence student performance, but simultaneously, it is a quantifier for other components within and outside the academic environment. These external elements are public services, private life or family support, which can affect the health, mental state or financial stability of a student, and indeed, his performance. Moreover, the student's achievement is affected by cognitive and non-cognitive attributes, such as discipline, planning abilities or intelligence [1]. Nonetheless, not only does this metric shows the ability of a student to comprehend and learn the proposed materials, but also the ability of the educational environment, to prepare an individual in the corresponding subject. Impactful components of the academic environment can be the teaching staff, the library facilities or the accessibility to the necessary bibliography.

After the creation of the internet, universities started to consider what benefits this technology can bring to them. In the beginning, there were technologies such as wiki pages or online forums, and the institutions tried to showcase to the students the opportunities that were available to them. This new concept was sometimes met with reluctance by the main stakeholder of the educational system [2]. Nowadays, the usage of tools such as virtual learning environments (VLE), online learning platforms or learning management systems (LMS) is a norm in higher education.

These platforms allow the collection of massive amounts of data, which consisted in an impactful improvement for higher education. This data permitted new insights, which led to an increase in the engagement and the overall experience of the students, with the VLEs. This new type of collected data doesn't resume only to the grades, but also to the entire activity of the student in the VLE, from the number of clicks to the type of educational resource that has been accessed. This latest educational data facilitated the creation of new research groups [3]. These communities can be associated with learning analytics, academic analytics, or predictive analytics. When we focus on the predictive analytics category, the target of the researchers was to find potential students at risk, so that the necessary support could be offered. Another target of the researchers consisted in matching the students with their career paths, to increase retention in regard to the completion of their degrees [4].

Previous research named this new field Educational Data Mining [3]. Numerous methods were used to understand and extract the essential information from the data, from statistical analysis to machine learning (ML) [5]. A sub-category of ML is deep learning, and its benefits have been used in the past for predicting student assessment results, student achievement and handwriting recognition [3].

In this research we will focus on predicting student achievement, using 3 machine learning algorithms. How a student's achievement is perceived on a general level, depends on the location and system in which that undergraduate is part. When we refer to North America, the metric that shows the performance of a student is the Grade Point Average (GPA). In this instance, trying to predict this result, would consist of a regression problem. On the other hand, in Europe, the most relevant

assessment criteria are the European Credit Transfer and Accumulation System (ECTS) [1]. In this case, predicting the achievement can be reduced to a Pass-Fail problem, for the courses that a student is taking.

In well-developed countries, individuals that have a tertiary education obtain numerous opportunities in the labour market. These benefits consist of higher earnings than the people without a tertiary qualification and higher chances of being employed [6]. Despite all the progress and investments that were done in the academic environment, there are still aspects that affect a significant amount of stakeholders of the society. In 2019, Organization for Economic Cooperation and Development (OECD) presented that in the Netherlands, 12% of the students that started a degree were not part of tertiary education in the following year. Moreover, OECD mentioned that in 2017, only 28% of bachelor students finished their degree in its theoretical duration. This problem exists also outside of the Netherlands, considering that the average of OECD countries, regarding students that obtain their bachelor's in the expected time frame, is 39% [6]. These represent problems that affect not only the students but also the economy since this translates into fewer years that can be spent in the labour market.

1.1 Research Questions

To solve these problems, our research will focus on finding the answer to the following research questions:

Q1. Which factors have an impact on a student's achievement?

Q2. Is it possible to predict student achievement at the midpoint of a course?

The creation of accurate machine-learning models takes a long period of time and it is a process that starts with gathering the data and ends with the analysis of the results. To have these algorithms work in the best way, they need to be used and trained for specific cases, based on the structure of the course and the available data. By answering Q1, the researchers will be able to know what features should they focus on, so that the duration of the development of these algorithms is faster and more efficient.

During a module, the lecturer or the course coordinator can analyse the assessment results of the students, to see if there can be made improvements later in the class. The possible data that is given by a VLE can offer more insights about the real situation of the group and by answering Q2, we want to predict if the student will pass or fail the course, considering only the data from the first half of the module. In this way, the lecturer would be able to identify at an early stage if the students from a cohort need help, so that he can apply the corresponding changes, to increase the performance of the students. The midpoint of the course (50% of its duration) was chosen, since it represents the moment in which the most data is collected, in rapport with the amount of time remaining till the termination of a course.

1.2 Thesis Outline

The structure of the paper is summarised as follows: Section 2 presents a state of the art in the field of predictive analytics. Section 3 consists of the methods and techniques that were used. In the end, Section 4 has the results and conclusions of the research.

2 State of the art

The field of student achievement prediction is multidimensional and can be seen from numerous perspectives, using countless methods to extract essential information from several data sources [3]. The potential directions which were explored by predictive analytics can be categorised into classification and regression problems. For classification problems, past papers predict if a student will pass a module or not, after a specified amount of time in the course, or if the student will drop out at the end of the year. On another hand, examples of regression problems are predicting the grade in a future assessment or in the final exam.

The summary of the ”Literature review” can be seen in Table 1. In this table, there are mentioned only the papers that have as main task to predict student achievement.

2.1 Literature review

In 2019, Waheed et al. [3] used the Open University Learning Analytics dataset (OULAD), to predict at-risk students. The OULAD presents as a final result 4 categories: ”Pass”, ”Fail”, ”Distinction” and ”Withdrawn”. In this research, they create a binary classification problem, in which they consider 4 pairs of results, these being: ”Pass/Fail”, ”Distinction/Fail”, ”Distinction/Pass” and ”Withdrawn/Pass”. Their data mining focused on the clicks that the students were doing in the VLE, creating 54 features. As methods for these binary classification problems, they used Deep Artificial Neural Networks (ANN), Logistic Regression (LR) and Support Vector Machine (SVM). Their conclusion mentions that the student’s activity in the VLE and the demographics have a significant impact in predicting academic success. On the other hand, they mention the benefits of Deep Learning in the early prediction of the performance of the students.

In 2020, Domladovac [5] is making a comparison between Neural Networks (NN) and 4 machine learning algorithms to predict if a student will pass or fail a course. He utilized 313 observations, obtained from the result of merging 2 different datasets, one representing the log data from the Moodle LMS and grades of the corresponding course from the University of Zagreb, Croatia. For feature selection, he used the Pearson correlation coefficient, resulting in 28 variables. As methods, he used NN, Gradient boosting Decision Trees (GBDT), Random Forest (RF), LR and SVM. To increase the metrics of the models, he performed Grid Search for hyperparameters tunning and K-fold for cross-validation. He quantifies the results of the models considering recall and accuracy since he mentions the significance of reducing the False negative rate. The best metrics are 0.8741 recall and 85.94% accuracy, using GBDT.

In 2021, Riestra-González et al. [7] focused on the early prediction of student performance. A novelty of their paper is the amount of available data. They had access to the log data from the Moodle LMS of the University of Oviedo, Spain. It consisted of 5,112 courses and 29,602 enrolled students. This led to incredible diversity in the data, having all 3 types of education, on-campus, online and blended modules, multiple disciplines and types of degrees. Their target was to make the ML models as general as possible, mentioning the drop in the metrics of the models, when tested with new types of courses. Due to privacy reasons, they were not allowed to use grades, which directed to fewer available courses and the inability to use an important component in predicting academic performance. Because of this, they use the LMS assignments as the target variable for their predictions. The models were created considering the available data after 10%, 25%, 33% and 50% from the module time interval. They create a classification problem with 3 groups: at-risk, fail and excellent for which they

use 5 classifiers: CART Decision Trees (CARTDT), Naïve Bayes (NB), LR, Multilayer Perceptron (MLP) and SVM. To choose the best features for each model, out of the total of 63 features, they use recursive feature elimination (RFE). After 10% of the course, MLP gave the best results, whereas, after 25%, 33% and 50%, CARTDT and MLP showed the best metrics.

Naseem et al., in 2022 [8], present their research on predicting student attrition, focusing on undergraduates studying computing science. They have a binary classification problem, in which they use 5 machine learning algorithms, to predict if the student will drop out or not at the end of the academic year. They start by presenting that a bachelor's in computing science is attractive to numerous people, but an important number of accepted students will eventually never obtain their bachelor's degree in this field. The effects of this reality created a shortage of computer science specialists in the job market when digitalisation is an important component in the development of a country. Moreover, they mention that student achievement is an index of performance for an institution and an indicator for the government for making decisions about educational grants. In terms of methodology, they consider 3 stages: "Pre-enrollment", which consists of the period before the start of the course, "After the first semester" and "After the second semester". Based on these periods, they select the corresponding elements to be considered in the analysis. The data used is collected from a university in the South Pacific region, using Moodle LMS. It has 963 results and was collected between 2013-2017. Similar studies use 3 categories of features: demographics, assessments and online presence, whereas this research has a progressive element, which is the existence of the financial information. For their data preparation, they normalise it using the min-max method and they keep the rows that have missing values. On the other hand, due to unbalanced data, they use the Synthetic Minority Over-sampling Technique and for feature selection, they use the Boruta algorithm. In their prediction, they use DT, RF, NB, LR and K-Nearest Neighbour (KNN) classifiers, which were chosen based on the results obtained in previous research. As performance metrics for their models, they use accuracy, sensitivity, specificity and receiver operating characteristic (ROC) and area under the curve (AUC). The best results were obtained using RF and KNN in stages 1 and 2, whereas LR was the best in stage 3, with an accuracy of 86%. In terms of feature analysis, for pre-enrollment, the math grade of a student has great importance, whereas later in the academic year, the grades showed the most relevance in their predictions.

López-Zambrano et al. [9], present in 2020, a new perspective regarding student achievement prediction. They analyse if those machine learning models could be used in different contexts, focusing on the data received from the Moodle LMS. A considerable problem with these models is that they are dependent on the available data, in the specific context. On the other hand, in the educational ecosystem, the log entries and other information are not accessible from all the courses. This leads to the concept of "Portability", where the knowledge obtained in an environment can be used in a different setting. The searched aspect is if we have sufficient data, these models could be ported without losing performance. The used data represents 3235 students at the University of Cordoba, Spain, from 24 courses. As a methodology, they run 2 experiments, where either they group the courses based on the subject or by the number of activities in the LMS. The model used in this research is J48, and in both experiments, they train the model on a dataset of course and then test that model using the other modules within the group. The used metric is AUC and AUC loss and in general, there is a considerable drop in terms of the efficiency of the models, when tested with new data. In experiment 1, the best-obtained average is 0.68 and in experiment 2 is 0.58, which leads to the conclusion that portability can happen only under some circumstances. A limitation of the research is the interpretability of the decision tree of a model since the lecturers might not be able to benefit from them, due to differences in terms of the structure of the courses.

Tomasevic et al. [10], in 2020, analyse a binary classification and a regression problem, by predicting if a student is at-risk of failing the course and the final exam score. Moreover, they present the importance of the data and the ethical and legal aspects that need to be considered in such research, since they are dealing with demographic and assessment data. As a methodology, they decide which models to be used based on the 3 types of classification and regression techniques, those being ”Similarity-based approaches”, ”Model-based approaches” and ”Probabilistic approaches”. The chosen models are KNN with no weights and with distance weights, SVM with linear and radial basis function kernels, ANN, DT, NB and LR and the used dataset is OULAD. They had 3 categories of independent variables, demographics, with 3 features, VLE interactions, with 6 features and assessments, with 7 features. The size of the final dataset is 3166 entries and they run the models 7 times, having all the combinations of the 3 categories of variables. As metrics, they used F1 for class prediction problems and Root Mean Square Error (RMSE) for the estimation ones. The best result for the Pass-Fail of the final exam prediction was 0.96 F1 when using ANN and just the engagement and performance features and for the regression task, the most efficient model is ANN, with an RMSE of 14.59 after the last assessment. A method presented in the state of the art is the matrix factorization method, but due to the insufficient diversity regarding the course types of the OULA dataset, they mention it as future work.

Realinho et al. [11], in 2022, as part of the learning analytics tool created by the Polytechnic Institute of Portalegre, Portugal, predict the eventual dropout and achievement of the student. By doing this, they observe which are the independent variable with the most relevance. They introduce the problem in the context of Europe. They mention that Denmark, which has the highest completion rate of around 80%, is still not enough to not represent a problem for the economy. Moreover, for the progress of EDM, they present the necessity for extending the range of collected data by the universities, such as dropout and transfer motives. The used data presents an extraordinary diversity, containing new categories of data such as socioeconomic and macroeconomic aspects. The information is obtained from 4 sources, collected between 2008 and 2019 and led to 4,424 records and 35 attributes. Their task is a three-category classification problem, those being: ”Graduate”, ”Dropout” and ”Enrolled”. In the analysis is mentioned that the used data is unbalanced towards the ”Graduate” group and that they use Pearson Correlation to see the interconnection between the features. As a methodology, they use Permutation Feature Importance and as ML algorithms, they apply RF, Extreme Gradient Boosting, Light Gradient Boosting Machine (LightGBM) and CatBoost. In conclusion, they indicate that the most essential attributes are related to past performance.

Ljubobratović and Matetić [12], in 2020, predict the final grade that a student will obtain. On the other hand, they present the problems caused by the General Data Protection Regulation (GDPR) for researchers, in the context of EDM. A new key aspect that ML researchers need to take into consideration is the ”Right to explanation”, explaining in a non-technical approach, why and how the models operate. They mention 2 categories, ”Global” and ”Local” explanations. The first one refers to the how, but in some black box algorithms, this is nearly unworkable, whereas the second type refers to the factors that were taken into review. The dataset used has 408 entries, collected from one course and has 6 features. For database analysis, they use a data heatmap, observing that the ”Pass” class outnumbers considerably the ”Fail” one. They have a binary classification task, for which they use RF. They obtain a 96.3% accuracy and for interpreting the features of the model, they apply variable importance, Break Down model, Tree surrogate and Local Interpretable Model-agnostic Explanations. They conclude that the independent variables with the most relevance are labs and quizzes. The experts observe that to achieve better metrics, an analyst would use a black box algorithm, but in the real-world context, the aspect of interpretability needs to be taken into

consideration. Their future works consist in exploring new domains, to analyse the explainability of the models.

Alshabandar et al. [13], in 2022, focus on projecting student achievement in online courses. They present the concept of Massive Open Online Courses, mentioning the diversity of available materials and that students could perform better in such an environment. In this paper, they use 4,004 records from the OULAD and have 2 experiments. They create 2 types of behavioural features, where "Static" uses the data available from the first day of interaction with the VLE, until the end of the time interval. The second type of attribute is called "Dynamic" and it is calculated for a specific short interval. There are 2 kinds of problems. The first one is a 3-class classification, for "Pass". "Fail" or "Withdrawn", using features from the entire timeframe, without assessment-related independent variables. On the other hand, they have a regression problem, for predicting the assessments grade. There are 6-time intervals, corresponding to each assessment and for predicting a future grade, the previous achievement and dynamic attributes are used in the analyses. As metrics, for the regression problem, RMSE and R-squared were used, whereas for classification, accuracy, F1 and ROC-AUC were applied. As a methodology, they use RFE for feature selection and apply a train, validate and test split of the data. They utilise various ML techniques, those being RF, MLP with 2 hidden layers, NN with a single hidden layer, Gradient Boosting Machine (GBM) and Generalized Linear Model. For the regression problem, GBM and RF obtain the best results. For the classification, all the models obtain similar results, the best one being GBM, with 86,8% accuracy. They conclude that the most relevant feature for predicting the assessment grade is the previous assessment result. On the other hand, the interaction with digital resources has high importance when forecasting if a student will withdraw or succeed.

Zhang et al. [14], in 2022, in the context of engineering undergraduates in China, project student achievement, using tree-based ML algorithms. They present external factors that have an impact on academic performance, but they focus on the aspect of time in this situation. Their objective is early prediction and relates to the fact that the sooner the prediction, the more time there is to identify and help the students in need. The used data has 690 records, was collected between 2015 and 2017 and is obtained from the engineering department. This task consists of a 3-class categorization problem. They group the students into 3 classes, based on the GPA, applying predefined intervals. For the forecast, they use decision-tree rooted algorithms, because of their interpretability and the applied models are DT, RF and GBDT. In terms of features, they group the courses from the first semester into 6 groups, in the foretelling, using the average GPA of these groups, alongside the gender of the student. They use the data from the first semester of each of the 4 years, predicting the overall GPA. Since they focus on at-risk students, the metrics used are precision, recall and F1. The training-test split is 70% and 30%, using 100 Monte Carlo simulations, to obtain the final results of the models. The conclusion presents that the classifiers perform better when the prognosis is done by applying more data. On the other hand, the best outcomes after the first academic term are obtained using RF, obtaining 71.8% recall. In this case, hyperparameters tunning didn't improve the performance of the models. Future work consists of predicting the performance at the end of the second semester while using more features and different methods.

Table 1: Literature review summary

Id	Reference	Objective	Dataset	Methods	Metric	Value	Conclusion
[3]	Waheed et al. (2019)	Pass-Fail prediction	OULAD	ANN	Acc	88.62	VLE interactions and demographics are important
				SVM		85.65	
				LR		84.23	
[5]	Domladovac (2020)	Pass-Fail prediction	Moodle LMS	ANN	Acc	84.34	Ensembles have higher accuracy than simple models
				GBDT		85.94	
				RF		83.05	
				LR		84.65	
				SVM		83.67	
[7]	Riestra-González et al. (2021)	Early prediction (50%)	Moodle LMS	NB	AUC	88.09	VLE interactions essential for early prediction
				CARTDT		93.50	
				LR		93.43	
				MLP		94.66	
				SVM		93.38	
[8]	Naseem et al.(2022)	Early Drop-out prediction (End of first semester)	Moodle LMS	DT	AUC	68.85	Grades are essential for early prediction
				RF		72.52	
				NB		74.79	
				LR		75.23	
				KNN		61.16	
[10]	Tomasevic et al.(2020)	Pass-Fail prediction	OULAD	ANN	F1	96.45	ANN best results for classification and regression
				DT		95.07	
				LR		94.42	
[11]	Realinho et al.(2022)	Feature importance	Multiple sources	RF	-	-	Grades and socioeconomic features have highest importance
				XGBT		-	
				LGBT		-	
				CGBT		-	
[12]	Ljubobratović and Matetić (2020)	Pass-Fail prediction	Moodle LMS	RF	Acc	96.3	Labs and quizzes have the most impact
[13]	Alshabandar et al.(2022)	Pass-Fail-Withdraw prediction	OULAD	RF	Acc	85.4	Previous marks have high importance
				GBM		86.8	
				MLP		85.8	
[14]	Zhang et al.(2022)	Drop-out prediction	Grades and gender	DT	F1	82.24	Better results with more data
				GBDT		84.38	
				RF		87.31	

[a] ANN - Artificial Neural Network, SVM - Support Vector Machine, LR - Logistic Regression, GBDT - Gradient Boosted Decision Trees, CARTDT - CART Decision Trees, NB - Naïve Bayes, MLP - Multilayer Perceptron, DT - Decision Trees, KNN - K-Nearest Neighbour, XGBT - Extreme Gradient Boosting, LGBT - Light Gradient Boosting Machine, CGBT - Cat-Boost, GBM - Gradient Boosting Machine

[b] Acc - Accuracy, AUC - Area Under the Curve

3 Methodology

There are two research questions, these being Q1: ”Which factors have an impact on a student’s achievement?” and Q2: ”Is it possible to predict student achievement at the midpoint of a course?”. In this case, we need a corresponding dataset to be analysed, using techniques that already showed efficiency in Educational Data Mining (EDM).

Our approach for this research has multiple stages. Firstly, we decided which dataset to use, since the research is based around this aspect. After this, we understood and analysed the available data, making the decision of how we should design 2 subsets, to correspond to the specifications of Q1 and Q2. This step was followed by creating and extracting the features from the data collection and finally encoding and analysing the final dataset used in the research. All the actions related to the dataset are described in Section 3.1. At the end of this process, there are 2 subsets ”Q1_dataset” and ”Q2_dataset”. These 2 sets will go through 3 ML classifiers, these being Logistic Regression, Random Forest and LightGBM. We use the same performance metrics to quantify the performance of the models for both, Q1 and Q2, these being accuracy, F1 and Area Under the Curve. Specific to Q1, we need to assess the importance of the features. The process and the decisions related to the machine learning aspect of the research are presented in Section 3.2. A visualisation of the methodology of this research is presented in Figure 1

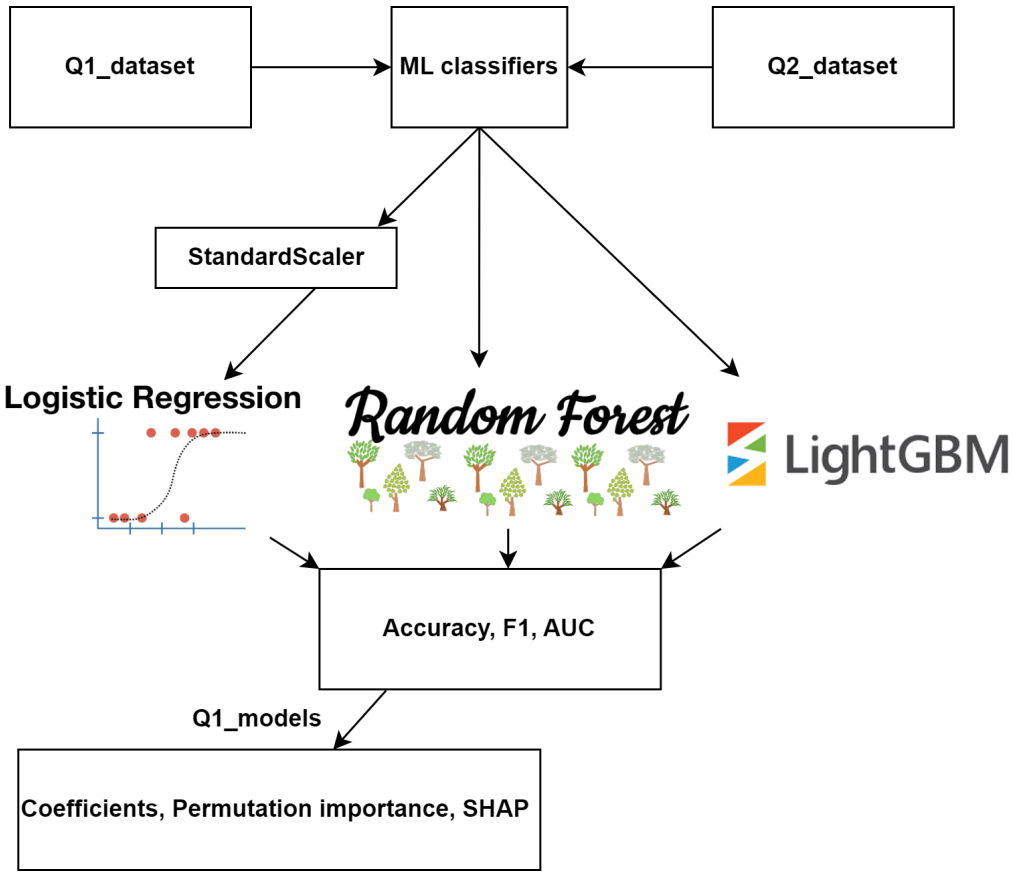

Figure 1: Methodology

There were numerous tools and platforms used in this process. For data analysis and machine learning, there was used Google Colaboratory, which is a Jupyter Notebook environment. The used Python

version was 3.10.12 and the additional used Python libraries were Pandas and Numpy, with versions 1.5.3, respectively 1.22.4. Logistic Regression and Random Forest were retrieved from scikit-learn version 1.2.2 and the version of LightGBM was 3.3.5. To store all the documents, datasets and the Jupyter notebooks, Google Drive was used, while when creating the documents, there was utilized Overleaf.

3.1 Dataset

In this research, the dataset represents the essence, since it is the primary source from which we want to extract meaningful knowledge, to be used later in the analysis. This section presents the arguments, ideas and concepts used, from deciding which dataset to use to the final dataset analysis part.

3.1.1 Dataset decision

In the initial stages of the project, we proposed to use the data collected by the Brightspace learning management system, from the courses held by the University of Groningen. By doing this, there would've been necessary numerous other stages that add difficulty and increase the duration of the research, without adding value to it. This is caused by the GDPR, since the data needed to be anonymised and to identify and eliminate the outliers from the dataset, to ensure the privacy of the participants. Besides this, Brighspace is used since 2022, which means that the current data is not sufficient for our research, considering we want to use machine learning models to answer the research questions.

In this case, we reoriented to use a public dataset, which meets our requirements. Sources of such datasets could be previous research from the learning analytics and predictive analytics fields, which either uses data from an LMS or their dataset (Table 1). Moreover, there exist platforms such as Kaggle or UC Irvine Machine Learning Repository which contain a large variety of datasets, sometimes their relevancy being questionable. From our analysis, we believed that the best dataset is Open University Learning Analytics dataset (OULAD) [15]. Its reliability is observed from the massive amount of research papers that used it, which allows us not only to understand it faster but also to compare our results with similar analyses (Table 1). This is a dataset easy to understand, yet complex enough to allow us to extract features related to demographics, interaction with the learning platform and the grades of the students.

OULAD is a dataset created by the Open University (OU) and released to be used for educational purposes. The data collection was launched in 2017, being administrated by Kuzilek et al. [15] and the data was collected in 2013 and 2014 from the VLE used by OU. A novelty of this dataset is that along with the demographic of the students, it contains the clickstream data of the undergraduates in the VLE. This data repository consists of 22 courses, from 4 STEM modules and 3 Social Sciences modules. The data repository has 32,593 students and over 10 million clickstreams. The raw data stream obtained from the VLE went through a process of cleaning and anonymisation. In this way, they increase the quality of the final data pool and respect the privacy of the students, the set being certified by Open Data Institute. In their privacy protection process, the ARX anonymisation tool was used, eliminating around 15% of the initial number of 38,239 students. This aspect represents the only disadvantage since it adds a certain degree of bias to the dataset, and indeed, to our research, since it alters the true representation of the real educational context from OU.

3.1.2 Dataset analysis

The dataset has 7 tables. How the tables are connected and what fields it contains can be found in Figure 3. In this set, there are 7 types of subjects, with could have multiple editions, happening either in 2013 or 2014, starting from February or October. The length of the courses varies, but the values are similar since the maximum length is 269 days and the minimum is 234 days. These courses present specific instances and it's hard to group multiple courses that have a different structure. All classes have at least one exam, whereas when we refer to the Tutor Marked Assessment (TMA) and Computer Marked Assessment (CMA), they vary from course to course and 2 times even between different editions of the same module. The structure of all editions of the 7 courses is represented in Figure 4.

Each student has a corresponding id, but from the 32,593 entries in the "studentInfo" table, there are only 28,785 unique ids. This is happening because the table contains the scenario of each entry in a course. In this way, there are students that not only took multiple different courses but also undergraduates that didn't pass a course and participated in a new edition. This leads to the observation that the actual id for each entry is a tuple containing the values from "code_module", "code_presentation" and "id_student". Furthermore, the age of the students is given as an interval, the options being 0-35, 35-55 and 55<=. On the other hand, the "final_result" column has 4 types of values: "Pass", "Fail", "Distinction" and "Withdrawn". In this research, we will reduce this column to a binary classification, where class "Distinction" will be considered a pass, while if we consider the data from the "Withdrawn" group, it will represent a fail.

The "studentAssessment" table represents the submitted assignments, containing aspects such as the submission date and if the assignment is banked from a previous course attempt or not. On the other hand, the score of each assessment is given, having values between 0 and 100. There are 173 entries that have no grade mentioned, the majority being of TMA type. Furthermore, the "studentVle" table contains the clickstreams that the students made in the VLE. The columns present each activity that happened, containing information about the date, the type of resource and the total amount of clicks.

3.1.3 Dataset design

In our case, there are 2 research questions, which require different answers. Because of this, we need to design for each of them, a dataset which will allow us to extract the necessary knowledge. The main source for both data collections will be OULAD, but the way in which we choose and query the data is specific to each question.

For Q1, we want to find what are the most important features that an ML algorithm takes into consideration when predicting the final result. In our analysis, we will consider all the scenarios available in the dataset, without the cases in which the student either withdrew, had no assessments submitted or had no VLE interactions. Besides these, there are no other query selectors, such as a specific timeframe. We use this approach because these classes of students have fewer assignments submitted and lower activity in the VLE, these being instances that the models would easily detect. These situations are categorised as outliers in our research. By using these examples in the training or testing phase, we would alter the results. What we want to find out, in this specific case, is which features predict an eventual pass or fail of a student that made progress towards passing the course. After choosing these instances, there are 20,285 results in the "Q1_dataset" and a graphic representation of the decisions can be seen in Figure 2.

For Q2, we want to see if it is possible to predict if a student will pass or fail the module from the midpoint of the course. This means that in the analysis, we will consider all the available information obtained from the first half of the course, including the data collected before the start of the module. The data is chosen when the date of that entry is smaller or equal to the length of the corresponding programme, divided by 2. On the other hand, the students that withdrew until the first 50% of the program won't be considered in the investigation, since their final result is already known at that moment. The students that withdrew after the half point, will be considered as "Fail". In the end, the "Q2_dataset" has 23,661 entries and an illustration of the querying process is in Figure 2.

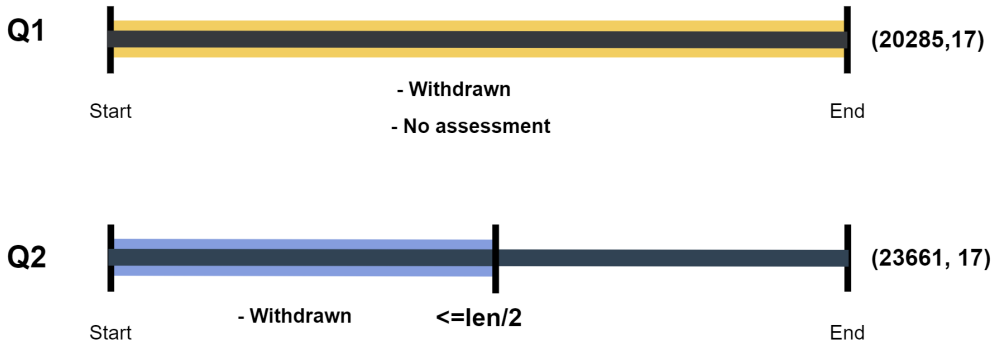

Figure 2: Dataset design for Q1 and Q2

3.1.4 Feature engineering

In this part, we will describe the reasons and the process by which we chose and created the features that will be used as input data for the ML models. When we relate to the OULAD, the extracted features can be classified into 3 groups, those being: demographics, assessments and VLE interactions. Moreover, when creating the independent variables, our target was to choose the ones that can be translated also into other contexts, without considering specific information to the Open University. In this way, the results and the methods used in this analysis would be more relevant also for other researchers. Nonetheless, the "Q1_dataset" and "Q2_dataset" are formed by the same set of variables, the creation of the features being done considering their corresponding design.

From the "studentInfo" table, we will extract the demographic features. There are 12 columns, 3 of them consisting of the identification tuple. In our analysis, we will use the "gender" feature, which is represented as "M" and "F" and "age_band", being given as 3 potential age intervals. Moreover, the set offers a feature related to the financial situation of the student, represented as the index of multiple deprivations, this one being "imd_band". On the other hand, we use all the features related to previous academic aspects, these independent variables being called: "highest_education", "num_of_prev_attempts" and "studied_credits". Lastly, the "disability" variable is used and it has 2 potential values, stating if the student declared a disability. To be mentioned that we don't use the "region" variable, since the values consist of regions only from the United Kingdom, in this way, we maintain the general aspect of our analysis. To conclude, these 7 features related to demographics, are selected from the table, without suffering modifications.

When creating the independent variables related to the assessments, we considered the presence of 22 editions of the 7 types of modules in the dataset. Since the structure of the courses differs, we created features such that we can extract knowledge from all of the courses. The features consist

either of general aspects that apply to a program or averages of the specific type of assessments. After we merge the ”assessments” table with the ”studentAssessment” table, we centralise all the necessary information for calculating the variables.

There are two types of assessments, these being Tutor Marked Assessment (TMA) and Computer Marked Assessment (CMA). There can be multiple assignments of each type during a single course, but their importance in the final grade is given by a given weight. To start with, we have two features in which we don't take into consideration the weight, these being ”TMA_average” and ”CMA_average”. These independent variables are obtained by calculating the total sum of the ratings and the total amount of graded assessments for each of the 2 types. Furthermore, we divide these 2 components, obtaining the mean of the TMA and CMA results. Similar to these features is the ”Total_average_score” attribute, which is the overall sum of the scores from all the assessments, divided by the total number of graded assignments. We use these features since we believe that a student who shows constant good results in all the assignments will have greater chances of passing and otherwise, we expect the opposite. Moreover, we have the ”TMA_score” and ”CMA_score” variables. These are calculated by summing the product of the score of an assignment, with the corresponding weight. In this way, we obtain the values that will be used in calculating the assignment result component from the final grade in the course. Moreover, we have the ”Total_assessment_score” feature, representing the final assessment grade, which represents the sum of the ”TMA_score” and ”CMA_score”. This group of features is highly correlated with each other, but we believe that a decisive predictor when forecasting future achievement is the situations when the student was previously marked. Lastly, there is the ”submission_date_average”, which is obtained by calculating the sum of the difference between the submission date of the homework and the deadline and dividing it by the overall amount of assessments. In this way, we try to understand if the behaviour related to the submission moment is important. This concludes with 7 features related to assessments and all of them were obtained by performing different calculations using the information from the dataset.

In the ”studentVle” table, there are stored all the interactions that the undergraduate had in the VLE, giving the date, the number of clicks and the type of resource that was accessed. The first VLE interaction feature is ”sum_click”, which represents the sum of all the clicks that a student made in the VLE. We created this independent variable since it is a relevant coefficient that shows the interest of the student towards the available resources. On the other hand, we have the ”consistency” variable, which represents how consistent is a student towards the interactions with the materials. To calculate it, we build a histogram to count the number of clicks for each day, from the day of the first VLE interaction, until the last one. To obtain the final feature, we calculate the standard deviation of the previously obtained array, The ”consistency” feature represents the novelty aspect of our variables list since previous research does not use a similar attribute in their research.

3.1.5 Feature encoding

In this analysis, we will be using multiple machine learning algorithms to answer the research questions. In general, these models can't use complex data types, such as images or strings. Because of this, the features need to be encoded to an understandable data type for the ML models, this being numerical data. The correct encoding of the data can lead to an improvement in terms of the performance given by these procedures.

We have a list of 16 features. From this list, 2 variables are given as strings, with 2 existing values each, these being ”gender” and ”disability”. These features will be binary encoded, being transformed

into 0 or 1. Moreover, there are 3 attributes given as strings, but which have a hierarchical order between their instances. The variables from this category are: ”highest_education”, ”imd_band” and ”age_band” and will be ordinal encoded. Their initial values will change into values from 0 to the number of unique values minus 1, considering their tiered structure. All the other features had initially numerical values and these don't require further encoding. Besides the 16 attributes, there is the ”final_result” column, which represents the target variable of our research. It has 4 potential values: ”Pass”, ”Distinction”, ”Withdrawn” and ”Fail”. This data will be binary encoded, where the first 2 observations will be transformed to 1 and the last 2 values to 0. In Table 2, there presented a summary of the feature list, used in this research.

Table 2: Features list

Category	Name	Values	Encoding
Demographics	gender	M and F	Binary Encoding
	age_band	3 ordered intervals	Ordinal Encoding
	imd_band	10 ordered intervals	Ordinal Encoding
	highest_education	5 ordered levels	Ordinal Encoding
	num_of_prev_attempts	Discrete numerical data	-
	studied_credits	Discrete numerical data	-
	disability”	Y and N	Binary encoding
Assessments	TMA_average	Continuous data	-
	CMA_average	Continuous data	-
	Total_average_score	Continuous data	-
	TMA_score	Continuous data	-
	CMA_score	Continuous data	-
	Total_assessment_score	Continuous data	-
	submission_date_average	Continuous data	-
VLE interaction	sum_click	Continuous data	-
	consistency	Continuous data	-

3.1.6 Final dataset analysis

There are 2 final datasets used in the analysis, these being ”Q1_dataset” and ”Q2_dataset”, being created according to the planned design. We will proceed into understanding them on a greater level, presenting the steps of the process, after which we obtained the 2 data collections. We will focus on the ”Q1_dataset” since both of them have similar values and distributions. To start with, we observe that there are 1,111 null values for the ”imd_band” column and the rows containing these values are dropped. After creating the features presented in section 3.1.4, we obtain a table with 6,664 students that had no assessments submitted. These rows will be deleted, after this step, remaining 48 students that have no VLE interactions. These ”Not a number” values will be changed to 0. Furthermore, the last step is dropping the rows where the ”final_result” equals ”Withdrawn”. After the encoding process presented in section 3.1.5, we obtain the final set, with no missing values. Moreover, the data collection has 53% males and 47% females and the ”highest_education” variable has a bell-shaped

distribution. In addition, the ”imd_band” has a uniform distribution and the ”age_band” has 69% of results in the 0-35 group. An important observation is that the mean of the ”TMA_score” and ”CMA_score” is 52.17, respectively 28.63. This shows that in general, the TMA assessments have a bigger weight in the final grade than the CMA ones. Lastly, the mean values for the ”TMA_average” and ”CMA_average” columns are 70.54 and 49.30, which shows that the students perform better when the assessments have greater importance.

3.2 Machine learning analysis

In EDM, numerous techniques have been used to obtain knowledge that could help a stakeholder in the field of education. Examples can be descriptive statistics and hypothesis testing [1] or machine learning and deep learning (Table 1). In our analysis, we will use machine learning algorithms, because they can be used for predicting and recognising patterns and due to their ability to handle multidimensionality, in both categorical and continuous data. Our task is a binary classification problem, where we have a medium-sized dataset, which presents categorical and numerical data and is essential to be able to get the importance of the features. Because of this, we opted for 3 models, those being Logistic Regression (LR), Random Forest (RF) and Support Vector Machine (SVM). This decision was based on the previously obtained results in similar research from the predictive analytics field, which were related to our situation. In the preliminary state of the research, we observed that SVM required a massive amount of computational power and memory while obtaining similar results to LR and worse than RF. After observing this and the performance of RF, we decided to change from SVM to another decision-tree-based algorithm, this one being LightGBM. Furthermore, we will describe the algorithms that are used in this research and the other techniques that were applied in the ML part of the project.

3.2.1 Machine learning models

In this section, we will present in more detail the 3 ML models used in the research. To start with, the Logistic Regression [16] was created in 1958 and despite its name, it is a binary classification algorithm. It uses the sigmoid function to give the probability, having as the threshold point 0.5. It is known for its simplicity and speed, while it also gives the importance of the independent variables. Numerous past papers used this algorithm, giving good results in the context of early drop-out prediction. Nonetheless, this algorithm requires the used data to be normalised, to increase its performance.

Random Forest [17] was created in 2001 and it is used for classification and regression tasks. During the training phase, it creates an independent decision tree, being trained on a subset of the initial data collection, using only a subgroup of the available features. After building a multitude of such trees, in the testing phase, it calculates the average of the obtained marks from all the decision trees, to get the final result. This makes it an algorithm that is robust to overfitting and versatile. Lastly, the features that are used at the top of the trees from the ensemble are considered to be more important.

Light Gradient Boosting Machine (LightGBM) [18] is an algorithm created by Microsoft and launched in 2017. This is a gradient-boosting framework, known for its speed and ability to handle complexity in large datasets. Moreover, it is a gradient-boosting decision tree algorithm, being a leaf-wise model, which develops the tree on the leaf that has the maximum delta loss. The efficiency of the model comes with some downgrades, these being related to the sensitivity to overfitting, it needs careful hyperparameters tuning and it is harder to interpret.

3.2.2 Machine learning related techniques

The training phase is the moment in which the models extract knowledge so that in the testing phase, they could perform the best. In the "Q1_dataset", we have a 1:2.6 ratio of the target variable and in "Q2_dataset", a 1:1.6 ratio. In both instances, the data imbalance is not big enough to require an over-sampling technique. Moreover, the sampling is randomly done, but to keep the consistency of the results, we use stratify. In this way, we have a similar distribution of the values of the target feature in the training and testing set.

To increase the metrics of the models, we decided to use multiple techniques. On one hand, RF and LightGBM need an optimal set of hyperparameters to perform ideally. To start with, hyperparameters tunning is essential for the achievement of a model and represents the usage of the best group of hyperparameters, those being set before the start of the training phase. To make this selection, there exist methods such as Random Search or Manual Tunning, but we opted for Grid Search, due to its capabilities. This method involves trying all the combinations of the given values of hyperparameters, returning the set that obtained the best value for the chosen metric. For Q1, our interest is to increase the accuracy, while for Q2, we want to boost the AUC metric. For LR, we analyse the C, penalty, solver and max_iter hyperparameters, for RF, we examine n_estimators, max_depth, min_samples_split, min_smaples_leaf and bootstrap. In regards to control variables of LightGBM, we study num_leaves, max_depth, min_data_in_leaf, learning_rate and n_estimators. In this process, we use stratified cross-validation with 3-folds, since this method is robust and helps observing potential overfitting. The values that were used in the search process are in Appendix C and the final set of hyperparameters used for the models from Q1 and Q2 can be found in Table 5.

To increase the performance for LR, we need to normalise the data. Moreover, we decided to use the Standard Scaler to standardise the data, since it ensures that the data collection has the same standard deviation. Nonetheless, this method is sensitive to outliers, an aspect which is not problematic in our instance. It is calculated as:

$$X' = \frac{X - \mu}{\sigma}$$

and X is the initial array, μ and σ is the mean, respectively the standard deviation of the values of the vector and X' is the final array.

3.2.3 Performance metrics

To assess the performance of the models, we need a set of metrics which define how the classifier performs in the testing phase. In this research, we will use accuracy, F1 and Area Under the Curve (AUC). Accuracy calculates the percentage of correct predictions over the total number of samples. F1 is a metric obtained from the harmonic mean of precision and recall. Moreover, AUC is used in classification tasks, showing the performance of the model at all classification thresholds. In general, these are the most used metrics and a reason for computing all of them is that it allows us to make comparisons easier with other researchers, and vice-versa. To be mentioned is that for Q1, we will focus on accuracy, while for Q2, we target the AUC.

3.2.4 Feature importance

For Q1, we want to understand what features are the most important. In our case, LR is a model which can easily have the features interpreted, while RF and LightGBM are more difficult. Our approach consists of 3 agnostic methods, those being the Model Feature Importance (MFI), Permutation

Importance (PI) and Shapley Additive explanations (SHAP). For MFI, in LR we will retrieve the coefficients of the model, while for RF and LightGBM we will use their built-in function for feature importance. PI is a method which involves shuffling individual features and observing the drop in the performance, taking into account the interactions of the independent variables. SHAP gives a contribution value to the features, representing the impact in the prediction. The analysis will consist in observing the obtained values and deciding which attributes are essential.

4 Results and Conclusions

In this section, we will present the results that were obtained in the process of answering the 2 research questions. In the first instance, we will observe the performance change after the usage of hyperparameters tuning, followed by the obtained metrics and feature analysis for Q1 and Q2.

4.1 Hyperparameters tuning results

The method of hyperparameters tuning was used to increase the performance of the models. The initial hypothesis is that models such as RF and LightGBM, require this method since these variables affect the way the model is doing the analysis. On the other hand, LR is a model which could benefit from this process, but it is seen as optional. To assess the change in the performance of the models, before doing the hyperparameters tunning, we run the models using the default values. These values are subtracted from the simulations in which we use the hyperparameters obtained and the difference in performance can be seen in Table 3. For Q1, the hyperparameters were chosen to increase the accuracy metric, whereas, for Q2, the target was to increase AUC. After this analysis, the RF and LightGBM models get a considerable improvement, the most significant one being +0.33 AUC for RF, in the Q1 experiment. Moreover, the highest improvement for LR is +0.02 accuracy in Q1. These 2 observations confirm our initial hypothesis and show the relevancy of this method for specific algorithms.

Zhang et al. [14] mention that the hyperparameters tuning method didn't improve the performance of the models. This is a possible anomaly, considering that in our case, the models that are decision-tree based had an increase in performance. On the flip side, the models they used are all decision-tree based: DT, RF and GBDT.

Table 3: The changes in metrics after hyperparameters tuning

Models	Q1(Acc)			Q2(AUC)		
	Acc	F1	AUC	Acc	F1	AUC
LR	+0.02	+0.01	0	0	0	0
RF	+0.24	+0.19	+0.33	-0.15	+0.09	+0.19
LightGBM	+0.19	+0.14	+0.16	+0.21	+0.18	+0.17

4.2 Q1 Analysis

The Q1 research mentions "Which factors have an impact on a student's achievement?". From here, we will do an analysis of the obtained performance metrics and observe which features were considered more important by the models. In both instances, we will be doing comparisons with similar research.

4.2.1 Perfromance metrics

In our case, we have a binary classification problem, in which we predict if the student will pass or fail the course, without considering the final grade in the exam. In this way, we want to see what features are important in the prediction of student performance. For this research question, we focus

on accuracy, but the metrics weren't essential in this case. At the same time, the higher obtained metrics, the better our feature engineering and analysis are, showing how able we were in capturing the knowledge available in the dataset. In this analysis, we have an 80%/20% train-test split for Q1, while for Q2, we use a 70%/30% train-test split.

In this analysis, RF and LightGBM obtained similar results, whereas LR performed significantly worse. In all 3 metrics, LightGBM obtained the best performance, having an accuracy of 89.25, F1 of 92.85 and AUC of 93.94. In table 4, there are presented all the performance metrics obtained.

In our literature review, there are 5 papers that solve a similar problem, but we will focus on making the comparison with papers that use the OULAD dataset. Waheed et al. [3], obtained the best results using ANN, having an accuracy of 88.62. To be mentioned is that they present very complex research, having numerous tasks and choosing from a set of 54 features. In our case, RF and LightGBM perform slightly better than the results obtained with ANN. Domladovac [5] is making a comparison between ANN and other ML algorithms, in the context of predicting student achievement. In his case, the best accuracy was 85.94, obtained using GBDT. Out of all the research mentioned included in the literature review, the best metrics were obtained by Tomasevic et al. [10]. They use OULAD and by using ANN, they obtain an outstanding 96.45 F1. To be mentioned is that this metric was obtained by considering only features related to assessments and VLE interactions. In our case, we obtain an F1 of 92.85 using LightGBM, showing a considerably lower result. Alshabandar et al. [13], in 2022, obtained the most recent results from our literature review, using the OULAD. They use a diverse set of models and the best metric is 86.8 accuracy, being obtained with GBM. Even if our RF and LightGBM obtained considerably better results, Alshabandar et al. had a classification task with 3 groups, pass, fail and withdraw.

Table 4: Performance metrics of the models

Models	Q1			Q2		
	Acc	F1	AUC	Acc	F1	AUC
LR	84.39	89.49	89.40	77.32	82.59	84.32
RF	89.08	92.74	93.68	82.32	86.49	88.88
LightGBM	89.25	92.85	93.94	82.34	86.42	88.72

4.2.2 Feature analysis

The essence of Q1 is to find which features or groups of independent variables have the highest importance. To assess this, we used the feature importance, the permutation importance and SHAP. The name of the features and the corresponding value, for each of the 3 models and methods, is available in Table 6, for LightGBM, Table 7, for Logistic Regression and Table 8, for Random Forest.

In LightGBM, we can observe that the most important feature is the Total_assessment_score, which is the value of the component that will be used in calculating the final grade. After this, sum_click is seen as the most important feature by the built-in function of LightGBM for feature importance. Moreover, the other features related to the assessments are the most relevant, sometimes observing the presence of VLE interaction features, such as consistency. No demographic features were present in the top 6 most important features.

In Logistic Regression, the most important feature is Total_average_score, which is the mean of all the grades from the assessments, without considering the weight of the assignments. It is closely followed in relevancy by the sum_click feature. We start observing a pattern, by having again numerous assessment-related independent variables being seen with high importance. Consistency is a feature that is considered as important in all the 3 methods, representing how constant is the student in relation to the VLE interactions. Specific to this model, there are present 2 demographic features, these being gender and studied_credits.

In Random Forest, in all of the 3 methods, we have the same 2 features as being seen the most important, those being Total_assessment_score and TMA_score. TMA_score represents the used component of the final grade, obtained from TMA assessments. In this model, the only feature that is considered relevant and is not part of the assessments group is sum_click.

We conclude that the most important feature in the context of our analysis is Total_assessment_score. This is something that we expected, considering this is the most important available feature that shows the past performance of a student. This happens in two of our models, RF and LightGBM, these being the models that obtained the best performance. On a second tier of importance, there are 3 independent variables, these being TMA_score, sum_click and Total_average_score. This concludes that the more active a student is in the VLE, the higher chance he has to pass the course. This is seen from the Logistic Regression model, which expects a linear relation between the variables and the target variable. On the other hand, Total_average_score is highly correlated with Total_assessment_score, while in general, TMA_score is considered more important than CMA_score, since the weight of TMA assessments is higher on average.

4.3 Q2 Analysis

The second research question is ”Is it possible to predict student achievement at the midpoint of a course?”. We will do an analysis of the obtained performance metrics and compare it with research that made their prediction at 50% of the duration of the course.

In Q2, we focused on the AUC metric. Moreover, LightGBM had the best accuracy, 82.34, while RF had the highest F1 and AUC, 86.42 and 88.72. In table 4, there are presented all the values of the obtained metrics. In this regard, Riestra-González et al. [7], in 2021, predicted the student achievement in multiple stages of the course, including 50%. They have a classification task with 3 groups, obtaining an amazing result of 94.66 AUC while using MLP. This is a massive difference compared to our results. To be mentioned that a novelty of their paper is the amount of available data, being the most diverse dataset from all the papers in the literature review.

4.4 Conclusion

The answer to the first research question is that the factors with the highest impact on the student’s achievement are the ones related to grades. This concludes that when predicting future achievement, the most relevant aspect is past achievement. In the literature review, this is a conclusion obtained by most. Moreover, even if Waheed et al. [3], mentioned the importance of the demographics, our conclusion is that the features related to the demographics are the least important. After the feature importance analysis, we conclude that the features from the assessments group are by far the most important, while the ones from the VLE group are the second most essential. Considering how the ML models in the predictive analytics field are made and that it is hard to reuse them in other

environments [9], this perspective brings new potential to the field. Considering the GDPR, the usage of demographic data is the most difficult. If we are able to obtain good results without this category, this leads to a faster, yet smaller set of features that need to be taken into account, while still obtaining good performance.

For the second research question, our answer is yes, it is possible to predict future achievement at the midpoint of a course, with 88.88 AUC. Similar research shows there is room for improvement, but even such a metric allows the course coordinators to obtain new insights about their students, at an early stage. Moreover, in terms of performance, there is a considerable difference between LR and the other 2 models. This can be explained by the fact that LR expects a linear relation between the features and the target variables. While we have some categorical features that can have a difficult relation with the target variable, this can be a reason for a worse performance. At the same time, it is unexpected that LR is the only model that found importance in demographic data, which can explain its inability to capture relevance from the other features. On the other hand, Ljubobratović and Matetić [12] present that a researcher needs to make a trade-off in terms of interpretability and efficiency. This idea applies to our case, since the 2 black box algorithms, which are known for handling data complexity, RF and LightGBM, perform considerably better than the simple LR.

4.5 Future work

OULAD was a great dataset when it was launched, but since 2014, the VLEs made significant progress. Future work should consider the usage of data from modern VLEs, focusing on the feature engineering process for variables related to the VLE interactions. In general, this data has complex relations, so the usage of decision-tree-based models is to be considered. Moreover, even if numerous features related to assessments are considered important, future work should consider shrinking this list. The variables from this group have a high correlation, which could mean that not the actual meaning of the data is understood.

Bibliography

[1] M. C. Schippers, D. Morisano, E. A. Locke, A. W. Scheepers, G. P. Latham, and E. M. de Jong, "Writing about personal goals and plans regardless of goal type boosts academic performance," *Contemporary Educational Psychology*, vol. 60, p. 101823, 1 2020.

[2] H. Li and J. Yu, "Learners' continuance participation intention of collaborative group project in virtual learning environment: an extended TAM perspective," *Journal of Data, Information and Management*, vol. 2, pp. 39–53, 3 2020.

[3] H. Waheed, S. U. Hassan, N. R. Aljohani, J. Hardman, S. Alelyani, and R. Nawaz, "Predicting academic performance of students from VLE big data using deep learning models," *Computers in Human Behavior*, vol. 104, 3 2020.

[4] P. J. Piety, D. T. Hickey, and M. J. Bishop, "Educational data sciences - Framing emergent practices for analytics of learning, organizations, and systems," in *ACM International Conference Proceeding Series*, pp. 193–202, Association for Computing Machinery, 2014.

[5] M. Domladovac, "Comparison of Neural Network with Gradient Boosted Trees, Random Forest, Logistic Regression and SVM in predicting student achievement," in *2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO)*, pp. 211–216, IEEE, 9 2021.

[6] *Education at a Glance 2019*. Education at a Glance, OECD, 9 2019.

[7] M. Riestra-González, M. d. P. Paule-Ruíz, and F. Ortin, "Massive LMS log data analysis for the early prediction of course-agnostic student performance," *Computers & Education*, vol. 163, p. 104108, 4 2021.

[8] M. Naseem, K. Chaudhary, and B. Sharma, "Predicting Freshmen Attrition in Computing Science using Data Mining," *Education and Information Technologies*, vol. 27, pp. 9587–9617, 8 2022.

[9] J. López-Zambrano, J. A. Lara, and C. Romero, "Towards portability of models for predicting students' final performance in university courses starting from moodle logs," *Applied Sciences (Switzerland)*, vol. 10, 1 2020.

[10] N. Tomasevic, N. Gvozdenovic, and S. Vranes, "An overview and comparison of supervised data mining techniques for student exam performance prediction," *Computers and Education*, vol. 143, 1 2020.

[11] V. Realinho, J. Machado, L. Baptista, and M. V. Martins, "Predicting Student Dropout and Academic Success," *Data*, vol. 7, 11 2022.

[12] D. Ljubobratović and M. Matetić, "Using LMS activity logs to predict student failure with random forest algorithm," in *INFuture2019: Knowledge in the Digital Age*, Faculty of Humanities and Social Sciences, University of Zagreb Department of Information and Communication Sciences, FF press, 2020.

[13] R. Alshabandar, A. Hussain, R. Keight, and W. Khan, "Students Performance Prediction in Online Courses Using Machine Learning Algorithms," in *Proceedings of the International Joint Conference on Neural Networks*, Institute of Electrical and Electronics Engineers Inc., 7 2020.

[14] W. Zhang, Y. Wang, and S. Wang, "Predicting academic performance using tree-based machine learning models: A case study of bachelor students in an engineering department in China," *Education and Information Technologies*, vol. 27, pp. 13051–13066, 11 2022.

[15] J. Kuzilek, M. Hlosta, and Z. Zdrahal, "Data Descriptor: Open University Learning Analytics dataset," *Scientific Data*, vol. 4, 11 2017.

[16] D. R. Cox, "The Regression Analysis of Binary Sequences," Tech. Rep. 2, 1958.

[17] L. Breiman, "Random Forests," tech. rep., 2001.

[18] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," tech. rep., 2017.

Appendices

A OULAD schema

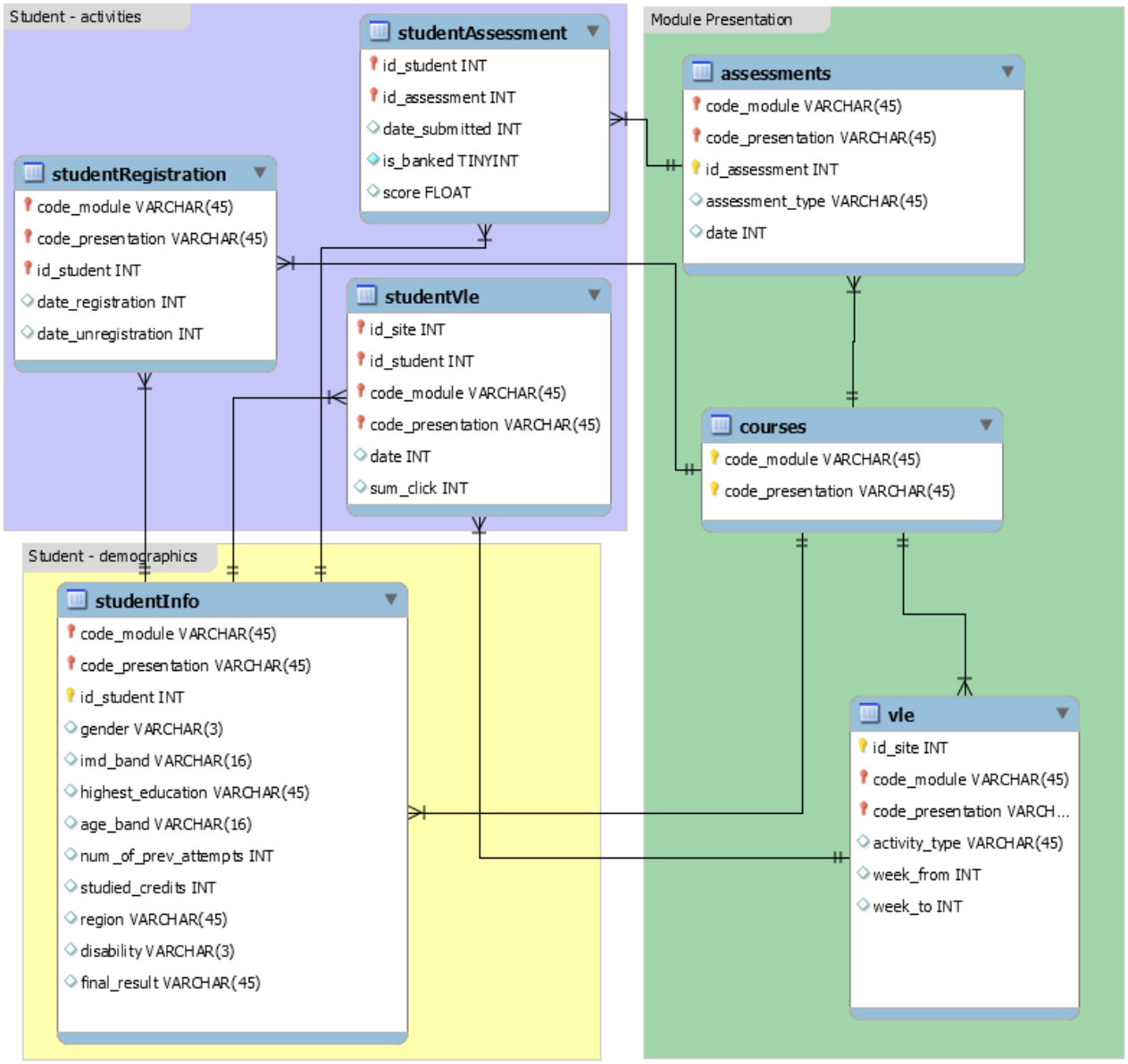

Figure 3: Dataset schema of the OULAD[1].

[1]Source: `https://analyse.kmi.open.ac.uk/open_dataset`

B Course structure

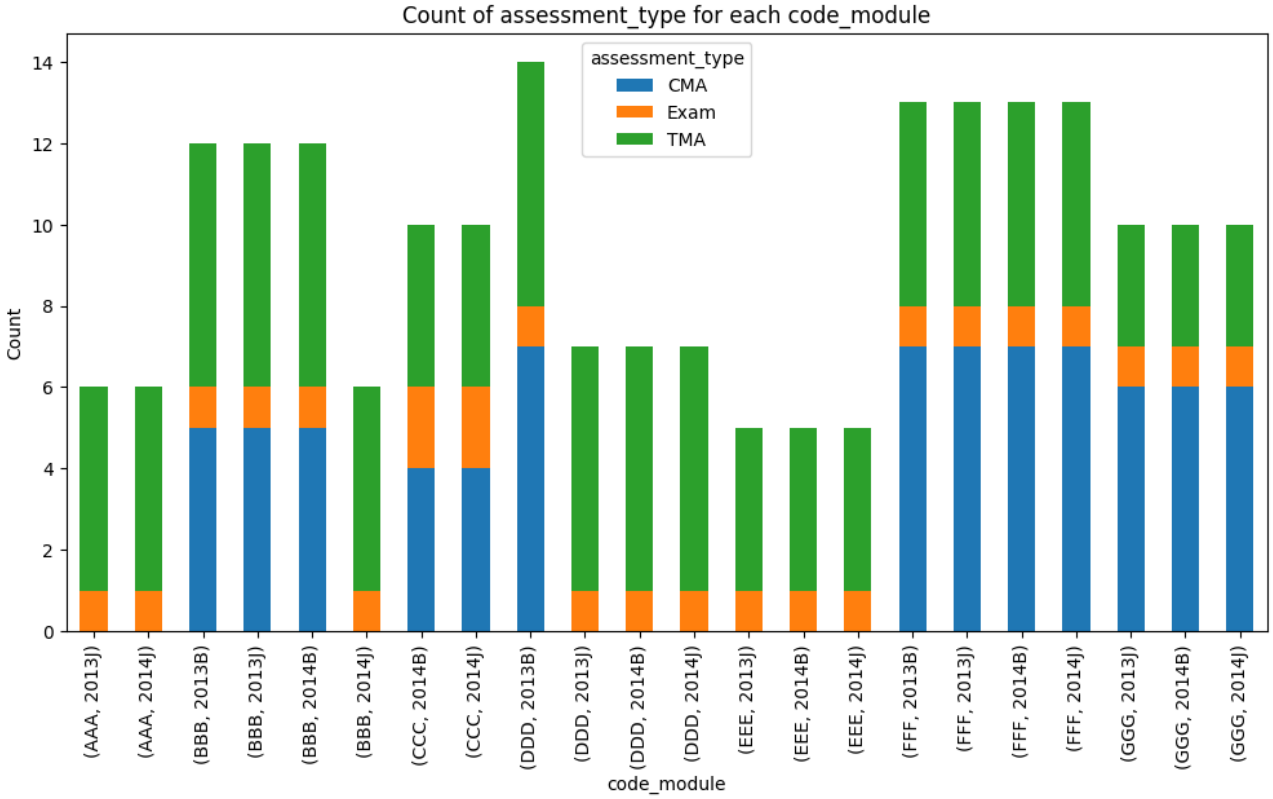

Figure 4: Course structure for all 22 editions

C Hyperparameters tuning values

The grid had the same values for Q1 and Q2. These are:

Logistic Regression:

- 'C': [0.001, 0.01, 0.1, 1, 10, 100]
- 'penalty': ['l1', 'l2']
- 'solver': ['liblinear', 'saga']
- 'max_iter': [10, 50, 100, 300, 500, 1000, 5000]

Random Forest:

- 'n_estimators': [100, 200],
- 'max_depth': [10, 20, 30, None],
- 'min_samples_split': [2, 5, 10],
- 'min_samples_leaf': [1, 2, 4],
- 'criterion': ['gini', 'entropy'],

LightGBM:

- 'num_leaves': [31, 127],
- 'max_depth': [5, 7, 9, -1],
- 'min_data_in_leaf': [30, 40, 50],
- 'learning_rate': [0.01, 0.05, 0.1],
- 'n_estimators': [100, 500, 1000]

D Hyperparameters final set

Table 5: The values of the hyperparameters used in the analysis

Question	Model	Name	Values
Q1	Logistic Regression	C	100
		max_iter	10
		penalty	l2
		solver	liblinear
	Random Forest	n_estimators	100
		max_depth	None
		min_samples_split	5
		min_samples_leaf	4
		criterion	entropy
	LightGBM	num_leaves	30
		max_depth	5
		min_data_in_leaf	40
		learning_rate	0.01
		n_estimators	1000
Q2	Logistic Regression	C	10
		max_iter	10
		penalty	l2
		solver	liblinear
	Random Forest	n_estimators	200
		max_features	sqrt
		max_depth	10
		min_samples_split	10
		min_samples_leaf	4
	LightGBM	num_leaves	127
		max_depth	7
		min_data_in_leaf	40
		learning_rate	0.01
		n_estimators	500

E Feature importance results

Table 6: Feature importance analysis for LightGBM

Pos	LightGBM					
	Name	Feat	Name	Perm	Name	SHAP
1	sum_click	408	Total_assessment_score	0.258	Total_assessment_score	1.982
2	Total_assessment_score	334	TMA_score	0.030	sum_click	0.554
3	submission_date_average	326	CMA_average	0.026	TMA_score	0.377
4	consistency	325	CMA_score	0.020	consistency	0.278
5	Total_average_score	323	sum_click	0.017	Total_average_score	0.254
6	CMA_average	281	Total_average_score	0.007	CMA_score	0.235

Table 7: Feature importance analysis for Logistic Regresison

Pos	Logistic Regression					
	Name	Feat	Name	Perm	Name	SHAP
1	sum_click	0.988	Total_average_score	0.050	sum_click	0.715
2	Total_average_score	0.879	Total_assessment_score	0.034	Total_average_score	0.689
3	Total_assessment_score	0.706	TMA_score	0.032	Total_assessment_score	0.600
4	TMA_score	0.701	sum_click	0.025	TMA_score	0.596
5	consistency	-0.49	studied_credits	0.016	consistency	0.344
6	TMA_average	-0.33	consistency	0.015	gender	0.306

Table 8: Feature importance analysis for Random Forest

Pos	Random Forest					
	Name	Feat	Name	Perm	Name	SHAP
1	Total_assessment_score	0.266	Total_assessment_score	0.099	Total_assessment_score	0.117
2	TMA_score	0.227	TMA_score	0.025	TMA_score	0.092
3	sum_click	0.102	CMA_average	0.024	sum_click	0.046
4	Total_average_score	0.073	CMA_score	0.011	Total_average_score	0.039
5	CMA_average	0.062	sum_click	0.009	CMA_average	0.026
6	TMA_average	0.061	submission_date_average	0.006	CMA_score	0.022