Effective Learning Resources to Improve Academic Performance: Insights and Recommendations

By

Sylvia Pereira

Course: Big Data Analytics Project
Supervisor: Professor Tamer Abdou

Abstract

The way learners interact and learn in a virtual environment continues to be an object of interest among researchers. Especially after the pandemic, many in-person courses are now 100% web-based, probably because this learning modality provides a convenient alternative for students with constraints and working professionals to learn on demand. According to a chart presented by the World Economic Forum, 189 million learners were enrolled in any course in 2021. The need for online learning on Coursera continues to outpace the pre-pandemic level, and this platform's registered learners increased exponentially from 44 million in 2019 to 92 million in 2021.

The main objective of this project is to leverage predictive learning analytics to provide insights into learners' behaviour within the Virtual Learning Environment (VLE). By analyzing historical data and identifying patterns and trends, academic institutions and companies can predict how learners will likely access resources in future learning experiences and what resources can be more effective in helping them pass a course. These insights can enhance students' learning journeys by offering meaningful resources they are more likely to access regularly. Ultimately, this project aims to improve the overall effectiveness of learning within the VLE and support learners' success.

This project considers the following research questions:

Question 1: Is it possible to identify the influential predictors that are most important in web-based learning?

Question 2: Which machine learning classifier offers optimal performance in predicting student results (pass or fail)?

Influential predictors can be understood as the available resources in the VLE, which comprehend but are not limited to HTML pages, PDF files, questionnaires, quizzes, collaborative activities, wiki pages, glossaries, forums, etc. Students have access to the materials, and their interactions with them are recorded (sum of clicks). This information is instrumental to tracking what resources learners typically access while taking a web-based course and which have proven to be the most efficient to help learners pass a course. A total of 4,689 sample students with two classes, pass and fail, and their click behaviour data from the OULAD (Open University Learning Analytics Dataset) is used. Although the full dataset presents 32,593 records distributed in seven unique courses, this project will focus on the BBB course for simplicity and model accuracy.

To answer the first question, association analysis is employed. *Content* and *subpage* appear six times among the first ten rules, meaning they are frequent items. The first rule, for example, shows that the support for the antecedent is 88%, meaning that 88% of the students accessed all three activities *quiz*, *subpage*, and *resource*. The support for the consequent is 71%, meaning that 71% of the students accessed all three activities, *forum*, *URL*, and *content*. The support for the combined antecedent and consequent is 67%, meaning that 67% of the students accessed all the activities in both sets. It's important to highlight that the association rules generated by the *Apriori* algorithm can provide insights into the relationships between features or items in a dataset. Still, they do not necessarily provide information about influential predictors of a target variable or how specific categories within features relate to the target variable. In summary, the *Apriori* algorithm offers insights into the relationship between different features but does not tell which class inside each feature can imply a course's passing or failing. By identifying the most

accessed resources, we can make future recommendations to learners and increase their performance in web-based courses.

To answer the second question, three different algorithms are applied to the dataset to predict students' outcomes – *Random Forest, Gradient Boosted and K-Nearest Neighbors*. The results demonstrate that *Random Forest* performed better accuracy, recall and F1-score, while *Gradient Boosted* performed better precision. Obtained results show that the feature selection step do not offer any improvements, so they are skip without affecting the models' prediction performance. *Random Forest* model without feature selection accomplishes 84% accuracy for the case of students passing a course.

Literature Review

The learning environment has changed drastically, especially after the pandemic in 2020. According to many articles from McKinsey, Harvard, and Forbes, the pandemic has forced schools, universities, and companies to remote work, which booms to the usage of virtual environments to learn and acquire new skills. Learners who used to take classes in person started to take online courses, and this study modality has proven efficient in many ways. A study by Frontiers in Education (2021) focused on biosciences courses in higher education reveals that "a majority of students reported positive experiences of online open-book assessments, and most would welcome this format in the future". Though the increase in online education is undeniable, and many institutions have invested time and money in providing learners with courses and resources to upskill, there are still big questions to be answered.

Question 1: Is it possible to identify the influential predictors that are most important in web-based learning?

Question 2: Which machine learning classifier offers optimal performance in predicting student results (pass or fail)?

This project will support SaaS companies and academic institutions to understand what and when resources in the virtual learning environment are accessed, which could help them drive investments in online programs. Although this study is not conducted to analyze data from a specific company, the data obtained is relevant and will provide enterprises with insights and ideas to be implemented in their future learning programs.

A virtual learning environment (VLE) is a designated information space not restricted to distance education, and it allows multiple technologies to be used and integrated into one system (Bashshar, 2017). VLEs in education describe a wide array of technological platforms that aim to

extend learning beyond physical space and provide interactivity between teacher and student or between student and content (Ketelhut & Nelson, 2021). A series of case studies published in 2013 by the Office for Standards in Education, Children's Services and Skills (Ofsted) described eleven examples of effective practices in VLEs. In general, a VLE is not only for storing resources such as PowerPoint presentations, handouts, animations, recorded lectures, and podcasts; it has many other functions to enhance teaching and learning, such as online tests, forums for discussions, and interactive quizzes. Also, the VLE is a vital element of learning strategy in many institutions, forming part of the overall strategic framework. Multi-dimensional VLEs allow learners to take different paths to learning and receive feedback and evaluations using multiple tools (Bashshar, 2017). A well-thought-out design, development, implementation, and evaluation strategy is needed for an effective VLE (Mueller & Strohmeier, 2011). As argued by Bashshar (2017), integrating web 2.0 tools such as wikis, podcasts, slide shares, broadcasts, and social networking sites into VLEs made them very potent for learners.

One of the main reasons academic institutions and companies invest in this environment is because students can learn based on their individual needs and because it offers a better cost-benefit since they can take any courses regardless of their geographical location. According to Mueller and Strohmeier (2010), these factors make VLEs ideal learning vehicles for corporate training. Technological functionalities, technology fit, and perceived usefulness by adult learners may influence how they interact and use VLEs. Research on perceived usefulness showed a high correlation between perceived usefulness and utilization (Mohr et al., 2012).

Related Works

Learning analysis has not been well-explored, but with the increasing use of technology as a fundamental learning resource, it has taken rapid shape over the last decade. More and more companies and academic institutions are interested in knowing how students behave in VLEs. According to Watershed, learning analytics involves the systematic measurement, collection, analysis, and reporting of data related to learners, learning experiences, and learning programs, intending to gain insights into how to enhance learning and improve learners' performance. Although vast data is available, it does not come in simple, well-organized, and collected formats. It exists in varied forms across systems and usually needs some transformation, so people can better understand students' learning journey.

One of the recent works using the same OULA dataset was a study conducted by Nikol Holicka (2020), a student from the Flatiron School of Data Science. She built a model that can help predict university students' outcomes and help the university to allocate additional support to those students who struggle. The course administrators could use her model to discover students likely to withdraw or fail and provide them with more support. This is not the primary goal of this project, but one interesting point of her research that aligns with my objective was the comparison among models. Nikol compared four classification models – *Baseline Random Forest* (60% accuracy), *Random Forest* with grid search (60% accuracy), *XGBoost* (72% accuracy), and *LightGBM* (73% accuracy). Ultimately, she validated *LightGBM* as the best-performing model with a 5% testing dataset; the final accuracy score was 74%.

Aljohani, Fayoumi and Hassan, in their article, tried to predict at-risk students using clickstream data in the VLE using the same OULA dataset. They "demonstrated that the clickstream data generated due to the students' interaction with the online learning platforms can

be evaluated at a week-wise granularity to improve the early prediction of at-risk students" (2019). The big highlight of their model is that it could predict *pass* and *fail* classes with around 90% accuracy within the first ten weeks of student interaction in a VLE. Their model offers a way for teachers to take an informed approach to advanced higher education decision-making toward sustainable education.

Another similar article was published in 2016 by Marbouti, Diefes-Dux and Madhavan, who also used predictive modelling methods to identify at-risk students early and inform both the instructors and the students. According to the authors, "the best model for overall accuracy was *K-Nearest Neighbor* (KNN), which identifies 94.9% of the students (failed and passed combined) correctly. This model was also the best for predicting the students who passed the course".

The research entitled "Predicting student success based on interactions with Virtual Learning Environment" (Doijode & Singh, 2016) is similar to this project. In this paper, the authors "predict students' success in an online course using regression, clustering and classification methods". One particularity of this paper is that the authors explore the whole dataset and consider region, age and gender as their study variables. Their work is inconclusive as they did not present the results of their models. The authors reinforce that the project scope will be extended to do a back test of the model and implement the successful validation results to identify at-risk students, applicable to online learning websites (Doijode & Singh, 2016).

The most recent research (related to the topic of this project) using the OULA dataset was published in 2021 by Josh Johnson. In his study "Predicting Student Success Using Virtual Learning Environment Interaction Statistic", the author describes the relationship between student behavior in the VLE and success. In this research, he does not consider students'

background, gender, disability, or socioeconomic situation. The goal of his project was "to model student interactions with an online learning environment to predict whether they will ultimately pass or fail a course early enough for intervention to be effective".

Methodology

A sequence of stages is performed before building the models. After combining the datasets in Alteryx and choosing the first group of features, I conducted a set of analyses with this newly prepared dataset to understand the correlation between the available features. The complete exploratory data analysis report (EDA) is found on GitHub.

The first model iterations included filter-based techniques but didn't produce good-performing models, so I decided not to use them. For example, when feature selection is used in a *Random Forest* model, the accuracy drops to only 55%. When using the wrapper-based selection, the resulting accuracy was better, around 78%, but still low compared to a model performed with all features. Some aspects can explain why feature selection is not working in this specific project. Looking at the correlation matrix (Figure 1), there are non-linear relationships between the variables, and filter-based techniques typically rely on simple statistical measures to assess the relevance of each feature. If the relationship between a feature and the target variable is non-linear, these measures may not capture the true importance of the feature and may overlook these features. So, this could be a possible explanation.

Another thing is that this dataset contains complex interactions. In some cases, the importance of a feature may depend on its interaction with other features. For example, *resource* and *quiz* features may not be significant on their own, but when combined, they become a strong

predictor of the target variable. Filter-based techniques may not be able to capture these complex interactions, and this can result in a poor model.

Fig. 1: Correlation Matrix

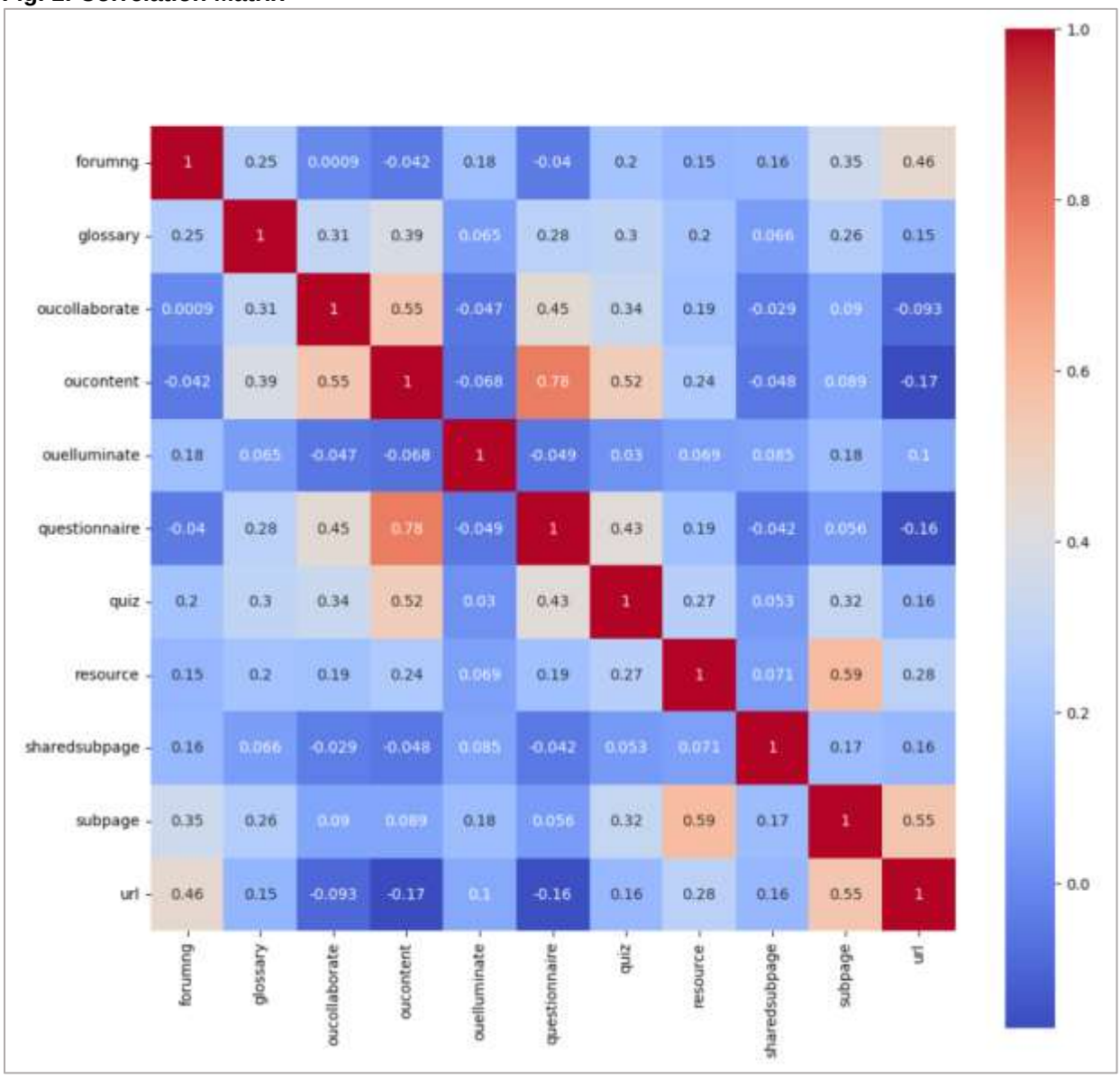
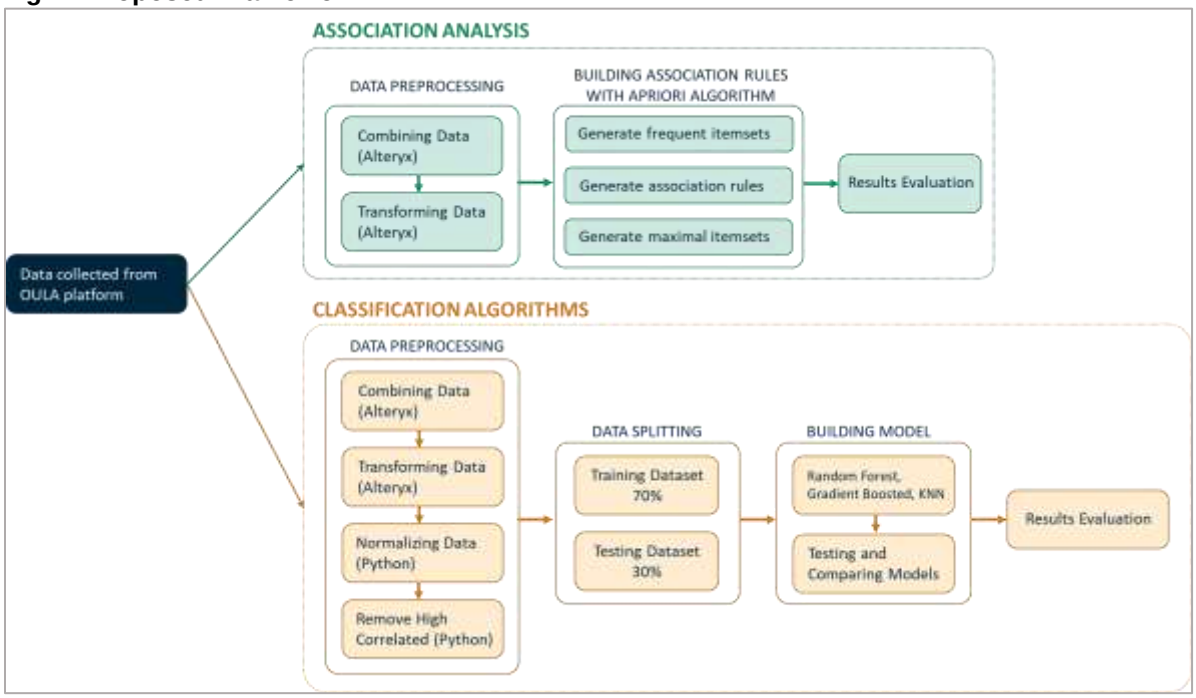


Figure 2 presents the proposed framework and outlines the main steps followed in analyzing both approaches, Association Analysis and Classification models.

Fig. 2: Proposed Framework



The Dataset

The dataset used in the project is from Open University, an online university that offers accredited degrees to distance learners. This dataset is divided into seven tables and contains information about seven modules (classes) from four semesters in 2013 and 2014. To meet the goals of this project, three files are used:

- **Studentinfo.csv**: This file contains demographic information about 32,593 students alongside their final results (pass, fail, distinction or withdrawn).
- **StudentVle.csv**: This file contains 10,665,280 records about each student's interactions with the materials in the VLE.
- **vle.csv**: This file contains 6,364 records about the available materials in the VLE (HTML, pages, pdf files, quizzes, etc.). Students access these materials online, and their interactions with them are recorded.

The original dataset contains regional, demographic, and personal data about students. Still, as mentioned before, this project's primary interest is finding the most effective learning resources to improve academic performance based on student behaviour while interacting with online resources in the virtual learning environment. In the studentVle.csv file available, the university provided information about each activity that each student interacted with, including the number of clicks on that activity.

The student interactions provided in the OULA dataset are filtered to consider only the BBB course. This decision was made because BBB has the highest number of passing students among the seven courses available.

Twelve different activity types are present in the BBB course. Each activity refers to a specific action, and the names are as follows*: forumng, glossary, homepage, oucollaborate, oucontent, ouelluminate, questionnaire, quiz, resource, sharedsubpage, subpage,* and *URL*. It's important to mention that, although feature selection was not applied considering a specific technique, the features *homepage* and *questionnaire* are removed from the dataset because they are high-correlated with other variables.

Categories and Features in the Dataset

The target of predictions in this project is the final result of a student in a specific course (pass or fail). Since the students with distinction grades passed the course, they are added to the "pass" class, resulting in Pass-Fail with 5,367 students (Pass: 3,752 – Fail: 1,615). The given students' academic performances comprehend a binary classification between students who

passed the course against those students who failed the course. Table 1 lists the virtual learning environment interaction features with a description.

Table 1: Description of predictive features in the dataset

Feature	Description	Type
id_student	Unique student ID	Categorical
final_result	The final result of each student in a given course	Categorical
forumng	Course total clicks on the discussion forum	Numeric
glossary	Course clicks on the basic glossary related to contents of course	Numeric
homepage	course total clicks on the course homepage	Numeric
oucollaborate	course clicks on the online video discussions	Numeric
oucontent	course total clicks on the contents of the assignment	Numeric
ouelluminate	course clicks on the online tutorial sessions	Numeric
questionnaire	course total clicks on the questionnaires related to course	Numeric
quiz	course total clicks on the course quiz	Numeric
resource	course total clicks on the pdf resources such as books	Numeric
sharedsubpage	course clicks on the shared information between courses and faculty	Numeric
subpage	course total clicks on the other sites enabled in the course	Numeric
URL	course clicks on the links to audio/video contents	Numeric

Data Preprocessing

Part 1: Alteryx Designer

Alteryx Designer Desktop is utilized to perform the preprocessing tasks. This phase may follow several steps: combination, data cleaning, reduction, and data transformation. Data cleaning is essential to remove unrelated data the model will not consider, and this step is done before starting to build the model in Python. Figure 3 provides an overview of the data preparation processes used in this project. A complete Exploratory Data Analysis report is available on GitHub.

Fig. 3: Data preparation in Alteryx Designer

Part 2: Python

The working dataset obtained from Alteryx Designer consists of 5,367 records. The second part of data preprocessing is performed in Python as follows.

Drop High-Correlated Features

After evaluating the correlation matrix, some collinearity is identified. In general, highly correlated features are not considered good predictors in a dataset because they do not add complementary information and can increase the risk of errors in the model created. If a predictive model has multiple predictors that are highly correlated, the algorithm has difficulty in trying to identify the proper coefficient to perform the analysis. From the 12 features in the working dataset, after trimming the high-correlated features (threshold > = 0.7), the number of new features is now 10. *Homepage* and *questionnaire* are dropped from the original dataset.

Normalize Data

Some features in the dataset present high variability. To address this issue, normalization rescales the values of numerical features to a standard scale, between 0 and 1 or -1 and 1, without distorting differences in the ranges of values. I applied normalization before feeding it into a

machine learning algorithm because it can help the algorithm converge faster and produce more accurate predictions.

K-Nearest Neighbors (KNN), a tested algorithm, use distance-based measures to determine the similarity between samples. If the features are not normalized, those with a larger scale or range of values will significantly impact the distance calculation, which can lead to biased results.

In this project, I introduce a model for predicting the students' results (pass or fail) based on student behaviour while interacting with the virtual learning environment. For the first part of the analysis, I used the *Apriori* association analysis to identify resources constantly accessed in the dataset and measure their support and confidence. By doing this analysis, companies and academic institutions might find that learners who access a *subpage* also tend to answer *quizzes* with high confidence, for example. This information helps recommend resources learners can access more efficiently during their learning journey. As mentioned before, the *homepage* feature was removed from the dataset because it is high-correlated with the *subpage*, and keeping both features could lead to inaccurate association analysis.

Table 2 shows the top five rules sorted by the lift metric in descending order. Minimum support is set as 0.4, meaning that only item sets that appear in at least 40% of all interactions in the dataset are considered for rule generation. The result for the first rule in the table shows that when users access the *quiz, resource, and subpage* together, they are very likely to also access the *forum, URL*, and *content* pages, with a confidence of 76%. The lift value of 1.07 indicates that the presence of the antecedent has a small positive effect on the occurrence of the consequent but not a particularly strong one.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
564	(resource, quiz, subpage)	(url, oucontent, forumng)	0.884802	0.711538	0.679238	0.767673	1.078891	0.049668	1.241616
577	(url, oucontent, forumng)	(resource, quiz, subpage)	0.711538	0.884802	0.679238	0.954605	1.078891	0.049668	2.537683
552	(url, oucontent, subpage, forumng)	(resource, quiz)	0.711165	0.885922	0.679238	0.955106	1.078093	0.049201	2.541063
589	(resource, quiz)	(url, oucontent, subpage, forumng)	0.885922	0.711165	0.679238	0.766702	1.078093	0.049201	1.238050
408	(resource, quiz)	(url, oucontent, forumng)	0.885922	0.711538	0.679238	0.766702	1.077527	0.048870	1.236450

Table 2: Top 5 rules generated by association rules

Some of the metrics used to analyze association rules include:

- **Support**: Shows the frequency of the itemset in the data.
- **Confidence**: Shows the probability of the consequent given the antecedent.
- **Lift**: Measures the strength of the association between the antecedent and consequent, considering the support of both. Values greater than 1 indicate a positive association.

Besides generating frequent itemsets, the *Apriori* algorithm is also used to extract maximal itemsets, which means an itemset that is frequent but not a subset of any other frequent itemsets. In other words, if an itemset is maximal, then no superset of that itemset is also frequent. For example, the itemset *subpage* has a support of 0.98, meaning it appears in almost all of the transactions in the dataset. Similarly, the itemset *resource* has a support of 0.96, indicating that it is also a commonly occurring item in the interactions. Table 3 shows the results of maximal itemsets.

Table 3: List of maximal itemsets using the *Apriori* algorithm

	support	itemsets
4	0.981143	(subpage)
3	0.968073	(resource)
18	0.963779	(resource, subpage)
0	0.911314	(forumng)
9	0.901979	(subpage, forumng)
2	0.900299	(quiz)
16	0.894884	(quiz, subpage)
8	0.892644	(resource, forumng)
28	0.890403	(resource, subpage, forumng)
15	0.885922	(quiz, resource)

In this project, three machine learning ML techniques are utilized and compared to see which classifier offers optimal performance in predicting if a student will pass or fail a course, *Random Forest*, *Gradient Boosted*, and *K-Nearest Neighbors*.

Model Building

Classification algorithms are essential for assigning students to different classes in educational data mining. Various algorithms can be used to handle this problem, including linear and radial support vector machines, artificial neural networks, decision trees, random forests, Naïve Bayes, Gaussian processes, and logistic regression.

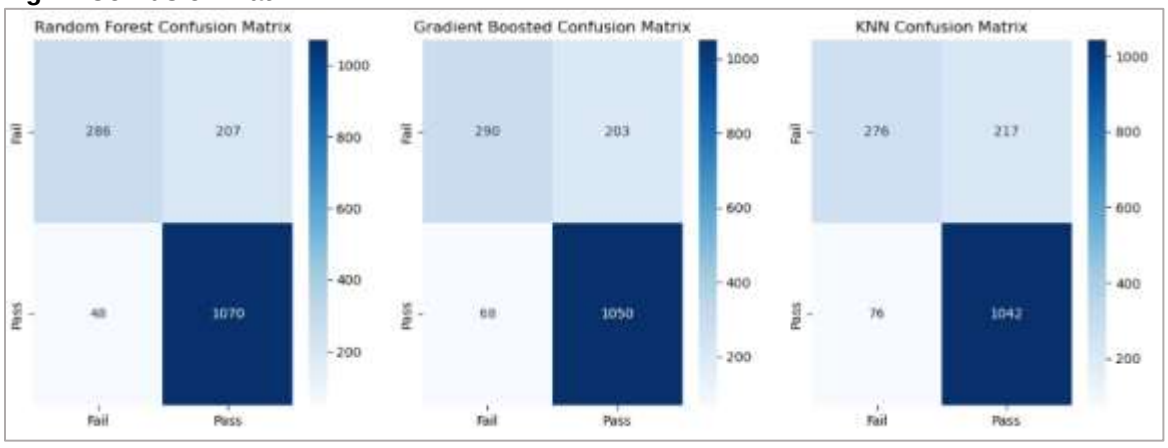
In this project, I utilized a set of classification algorithms that include *Random Forest Classifier* (RFC), *Gradient Boosted Classifier* (GBC) *K-Nearest Neighbors* algorithm (KNN).

The current study classifiers are trained using a dataset constructed from the VLE data. The input features used in training are the total number of clicks on the VLE activities completed in the VLE course BBB, and the target variable was the predicted student results (pass or fail).

I used a 10-fold cross-validation method to train and test the models. Cross-validation is primarily utilized to assess model performance. In k-fold cross-validation, the data is divided into k different subsets, the model is trained using k-1 subsets, and the remaining subset is used for testing. The average performance obtained from this method reasonably estimates model performance.

After training the models, I assessed the performance of the learning models using previously unseen data. I obtained the prediction results for the models with the test data and counted the number of true positives, true negatives, false positives, and false negatives that are used to evaluate performance. Through this process, I obtained the numbers of true positives (Pass) and true negatives (Fail) and the number of false positives and false negatives. Figure 4 shows the correlation matrix for all three models.

Fig. 4: Confusion Matrix



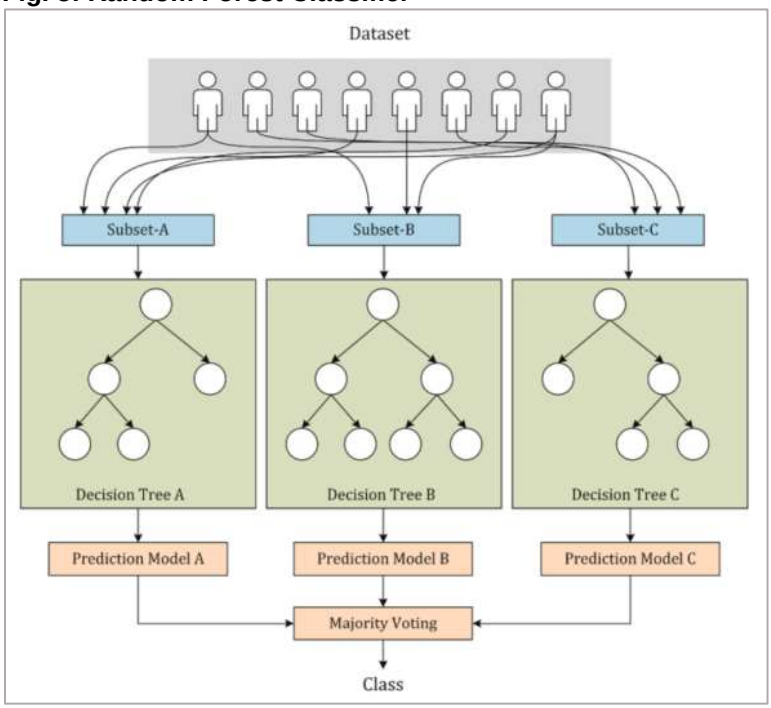
The *Random Forest Classifier* is a popular and powerful machine learning algorithm. I can infer that this model performed better in predicting students' final results because this algorithm is a robust model for dealing with outliers. Some of the features in the dataset present high variance. *Forum*, for example, has a minimum value of 0 and a maximum of 13,154. This classifier can also help identify the data's most important features. This model is an excellent example of ensemble learning where multiple weak learners are combined to create a stronger learner.

The algorithm is summarized in the following steps:

1. A random subset of students is selected from training data.
2. After removing the highly correlated features, a subset of features is chosen randomly.
3. A decision tree is constructed, and at each node, the feature that maximizes the information gain is selected to split the data into two subsets.
4. The splitting process continues until a stopping criterion is met. In *Random Forest Classifier*, the is the defined criteria:
    a. Number of trees = 100
    b. Maximum depth = 10
    c. Random state = 10
5. The steps above (1-4) repeat to create a forest of decision trees.
6. The final step is called voting. To predict a new data point, it is passed down to each tree in the forest, and the majority class prediction of the trees is taken as the final prediction.

Figure 5 helps to explain the working mechanism of the random forest algorithm for students' performance prediction (Computers and Education: Artificial Intelligence | Journal | ScienceDirect.com by Elsevier).

Fig. 5: Random Forest Classifier



Results

The results are presented in three main aspects: model training implementation, model performance (model evaluation results) and feature importance. The first two aspects show how the models are trained, how well they perform, and which model is the best. The third aspect, feature importance, examines the dominant features of the best model to gain insights into teaching and learning.

I use four commonly used metrics to measure the effectiveness of the models: Accuracy, Precision, Recall, and F-1 Score. Accuracy represents the proportion of the total number of correct predictions, while precision indicates how many values are actually positive out of all the

predicted positive values. Recall represents the proportion of correctly classified values out of all relevant values (correctly classified and unclassified), F-1 Score tries to find the balance between Precision and Recall by calculating their harmonic mean. These metrics provide a comprehensive understanding of the classifier's performance and can be used to compare different classifiers.

Table 4 provides the model performance for the test data, which forms 30% of the total number of student records; 1,611 students out of the total 5,367 are left for testing the model in each fold within the 10-fold cross-validation. Overall, the results show RFC > GBC > KNN. The best model is the *Random Forest Classifier* with 84.17% accuracy, 83.79 Precision, 95.70 Recall and 89.35 F1-Score.

Table 3: Model Performance

Metric	Random Forest	Gradient Boosted	K-Nearest Neighbors
Accuracy	84.35%	83.11%	82.68%
Precision	83.88%	84.11%	83.32%
Recall	95.88%	93.29%	93.28%
F1-Score	89.48%	88.46%	88.26%

Results for the *Random Forest Classifier* indicated that the classifier correctly classified 1,070 students out of 1,611 to the correct Pass class and 286 to the Fail class.

The important features of the best model, *Random Forest*, are examined by observing how the model's performance changes when the top 5 features in the model are eliminated. As a result, the model performance shows a dropped accuracy. The accuracy decreases significantly when the features *quiz*, *content*, *resource*, *subpage*, and *forum* are removed. Therefore, these five

features can be considered dominant. A comparison of the accuracy after the five most relevant features are removed is shown in Table 4.

Table 4: Feature importance analysis result

Removed Features	Original Model's Accuracy	Model's accuracy after removing features	Dropped Accuracy
Quiz, content, resource, subpage, forum	84.35%	73.92%	10.43%

The analysis demonstrates that students' clicks on *quiz, content, resource, subpage* and *forum* are the most important predictors of student engagement in the VLE course because these features appear most frequently in the classification model. Other features, such as the number of student clicks on *URL, collaborate, glossary, elluminate* and *sharedsubpage,* are less significant predictors of student success. This also proved true when analyzing the association analysis results, where the top five rules show precisely the same resources as antecedents or consequents (Table 2).

This study trained three predictive models using machine learning methods and clickstream data, achieving up to 84.34% accuracy. The results provide insights into practical ways to extract features, train and evaluate predictive models in student performance prediction tasks using students' clickstream data.

Conclusion

This project explores the potential of using clickstream data to predict student performance. Analyzing important features from the best model (*Random Forest Classifier*) shows that clicks on the *quiz*, *content*, *resource*, *subpage* and *forum* are significant in predicting student performance. Based on these findings, content developers, instructional designers, and educators can consider improving the online learning environment by utilizing the advantage of students visiting these resources more frequently.

The association analysis supports the results and is a critical evaluation technique alongside the classification model. In addition, although this project was conducted based on a data science approach rather than with an educational focus, the identified important features from the *Random Forest Classifier* can inform future course design and teaching interventions.

Some of the areas that can benefit from its results are:

- *Investments*: What resources should institutions invest time and money to develop?
- *New Course Developments*: Knowing that some resources are accessed more, should they focus on improving those or developing others? How could they maximize the use of the low accessed resources?
- *Predictions and Interventions*: Is building a learning performance predictor with high versatility and accuracy possible? Some institutions, especially in the academic field, can be benefited by creating a tool to predict the number of students that may pass or fail a course to decide what interventions they could do to increase the number of *passes* and decrease the number of *fails*.

Although this project is an important information source, it is impossible to infer conclusions about students' performance based solely on the number of times a student clicks

whenever they access the virtual learning environment. The results concern only one course and, therefore, cannot be generalized.

References:

Link to Github: https://github.com/sylviabpereira/Final_Project

(2013, January 1). *Virtual Learning Environments: Eleven Case Studies of Effective Practice*. Office for Standards in Education, Children's Services and Skills (Ofsted). Retrieved February 5, 2023, from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/383958/VLE_e-portfolio_-_case_studies_booklet.pdf

Aljohani, N. R., Fayoumi, A., & Hassan, S.-U. (2019). Predicting At-Risk Students Using Clickstream Data in the Virtual Learning Environment. *Sustainability*, *11*(24), 7238. https://doi.org/10.3390/su11247238

Bashir, A., Bashir, S., Rana, K., Lambert, P., & Vernallis, A. (2021). Post-COVID-19 Adaptations; the Shifts Towards Online Learning, Hybrid Course Delivery and the Implications for Biosciences Courses in the Higher Education Setting. *Frontiers in education*, *6*. https://doi.org/10.3389/feduc.2021.711619

Bashshar, C. E. (2017). *Virtual Learning Environments' Impact on Adult Learners' Motivation in the Workplace* [Doctoral dissertation, Walden University]. https://scholarworks.waldenu.edu/cgi/viewcontent.cgi?article=4487&context=dissertations

Diaz-Infante, N., Michael, L., Ram, S., & Ray, A. (2022, July 20). *Demand for online education is growing. Are providers ready?* McKinsey & Company. Retrieved February 5, 2023,

from https://www.mckinsey.com/industries/education/our-insights/demand-for-online-education-is-growing-are-providers-ready

Gallagher, S., & Palmer, J. (2020, September 29). *The Pandemic Pushed Universities Online. The Change Was Long Overdue.* Harvard Business Review. Retrieved February 5, 2023, from https://hbr.org/2020/09/the-pandemic-pushed-universities-online-the-change-was-long-overdue

Holicka, N. (2020, March 8). Predicting students' results with data from The Open University. Retrieved March 4, 2023, from Medium website: https://nikolh92.medium.com/predicting-students-results-with-data-from-the-open-university-997e21e9048e

Johnson, J. (2022, October 11). Predicting Student Success Using Virtual Learning Environment Interaction Statistics. Retrieved March 11, 2023, from GitHub website: https://github.com/Caellwyn/ou_student_predictions/blob/main/report/OU_student_predictor_presentation.pdf

Ketelhut, D. J., & Nelson, B. (2021). Virtual Learning Environments. *Oxford Bibliographies*. https://doi.org/10.1093/OBO/9780199756810-0288

Koksal, I. (2020, May 2). *The Rise Of Online Learning*. Forbes. Retrieved February 5, 2023, from https://www.forbes.com/sites/ilkerkoksal/2020/05/02/the-rise-of-online-learning/?sh=74d9b1a72f3c

Kuzilek J., Hlosta M., Zdrahal Z. *Open University Learning Analytics dataset* Sci. Data 4:170171 doi: 10.1038/sdata.2017.171 (2017). Retrieved January 22, 2023, from https://analyse.kmi.open.ac.uk/open_dataset

Miller, K. (2020, February 18). What is Learning Analytics & How Can it Be Used? Retrieved from Northeastern University Graduate Programs website: https://www.northeastern.edu/graduate/blog/learning-analytics/

Mohr, A. T, Holbrugge, D., & Berg, N. (2012). Learning style preferences and the perceived usefulness of e-learning. *Teaching in Higher Education 17*(3): 309-322, doi:10.1080/13562517.2011.640999

Mueller, D., & Strohmeier, S. (2010). Design Characteristics of Virtual Learning Environments: An Expert Study. *International Journal of Training and Development*, *14*(3), 209-222. https://doi.org/10.1111/j.1468-2419.2010.00353.x

SoLAR. (2019). What is Learning Analytics? - Society for Learning Analytics Research (SoLAR). Retrieved from Society for Learning Analytics Research (SoLAR) website: https://www.solaresearch.org/about/what-is-learning-analytics/

Wood, J. (2022, January 27). *These 3 charts show the global growth in online learning*. World Economic Forum. Retrieved January 22, 2023, from https://www.weforum.org/agenda/2022/01/online-learning-courses-reskill-skills-gap/