STAT642 FINAL PROJECT

ONLINE LEARNING

Neel Kumtakar

Jing Wang

Josh Pierce

Michael Sadi

1. Introduction

1.1 Background

Nowadays, in some countries, due to the development of the Coronavirus pandemic, classes in colleges and schools have transitioned into online formats. This unexpected change in teaching and learning environment can potentially result in an opportunity to harness the power of data and create better learning incentives for students and teachers via online learning. Proper and good education is necessary for everyone in order to improve knowledge, way of living as well as social and economic status throughout life. Online education is about anytime and anyplace access to the educational resources that all people need to enhance their lives and better understand the world around them.

1.2 Objective

This project aims at measuring and predicting the performance of students in online learning. This project results can help professors to identify some at-risk students (predict to fail) in the early stage, and provide them with academic help, like providing additional online office hours for these students. Also, the student can use the model to predict whether he or she would fail the course, then he or she could make good use of the time, like putting more time on the course if his or her predicted result is the “fail”. In addition, using the performance result (pass/fail) is one of the ways to evaluate whether students gain the knowledge the course delivered, and both professors and students could improve their teaching and learning process in advance based on the predicted result.

2. Data

2.1 Data Introduction

The original dataset is from (Kuzilek J., Hlosta M., Zdrahal Z. Open University Learning Analytics Dataset(OULAD) Sci. Data 4:170171 doi: 10.1038/sdata.2017.171 (2017). The

original dataset, consisting of 7 sub-datasets, contains the information about two domains (STEM and Social Science), 32,593 students, their assessment results, and logs of their interactions with the virtual learning engagement represented by daily summaries of student clicks (10,655,280 entries), the final result (withdrawn/fail/pass/distinction) over a course period of 9 months in 2013 and 2014.

2.2 Data Pre-processing

In this project, based on the original comprehensive dataset, following data preprocessing steps are taken.

- Data sampling. In the original dataset, some students took two domain courses (STEM or Social Science) in both 2013 and 2014, causing data repetition. In this study, only courses in 2014 will be analyzed. As such, courses in 2013 are removed. And the students who took more than one course in 2014 are deleted as well to make sure there is no repetitive data, and only contain unique student id. In addition, since the amount of STEM courses is significantly larger than that of Social Science, the Social Science courses are removed, and the STEM courses are extracted only
- Binary variable creation. In the original dataset, the *final_result*, as a dependent variable, consists of four types of value- distinction (which means top grade), pass, fail and withdrawn. Concerning the non-representativeness of students withdrawing course status, samples of "withdrawn" are deleted. Subsequently, among the three values left, "distinction" is merged into the "pass" category. As such, a binary dependent variable, only including "pass and fail" value is created, which is beneficial for classification and prediction in the following steps.
- Aggregation. In the original dataset, the "TMA_score (tutor marked assessment score)" and "CMA_socre (computer marked assessment score)" include students' scores of various quizzes during the term, respectively. Each test has a different score range, to reduce scores variability, all tests' scores are grouped by student ID using the function of "mean". As such, each observation will have one aggregated score with the same range (0-100).
- Feature creation. In the original dataset, data about the students' click time is distributed on both the different websites and online courses, rather than centralized. Here, the click time means that the times a student clicked a certain online course material for the entire course period. Since click time is an important metric that could be used to quantize students' interaction and engagement to that course. As a result, a new feature, *sum_click*,

is created. It combines all click times from different websites according to corresponding student ID using the function of "sum".

- Feature selection. The original dataset includes 20 online activity types, which means the website type chosen by one course to present its content, such as quiz, questionnaire and homepage, etc. Students learn online courses by clicking corresponding website types. However, some website types are selected by only 1 or 2 courses, lacking enough information. As a result, these online activity types are deleted, while other large amounts of website types are selected.
- Variable transformation. To better normalize the click times data distribution, and to map the entire set of click time value, log10(x) transform will be applied in the following exploratory data analysis.

2.3 Variables Description

Through pre-processing the original data, the final data used in this project is constructed. This data will focus on STEM-student online learning performance at the Open University in the U.K. in 2014. The dataset has 6,693 observations and 22 variables. Among these variables, there are 7 categorical variables and 15 numerical variables, which consist of students' demographic information, courses related information, three kinds of assessment scores, and click time data. For the click time data, it includes click time at 9 website activity types, with which a student clicks and interacts. In addition, it also includes the *sum_click* variables, which sums up all the click times that a student clicks a certain course material for the entire course period. More details of variables can be seen in *Appendix A - Variables description*.

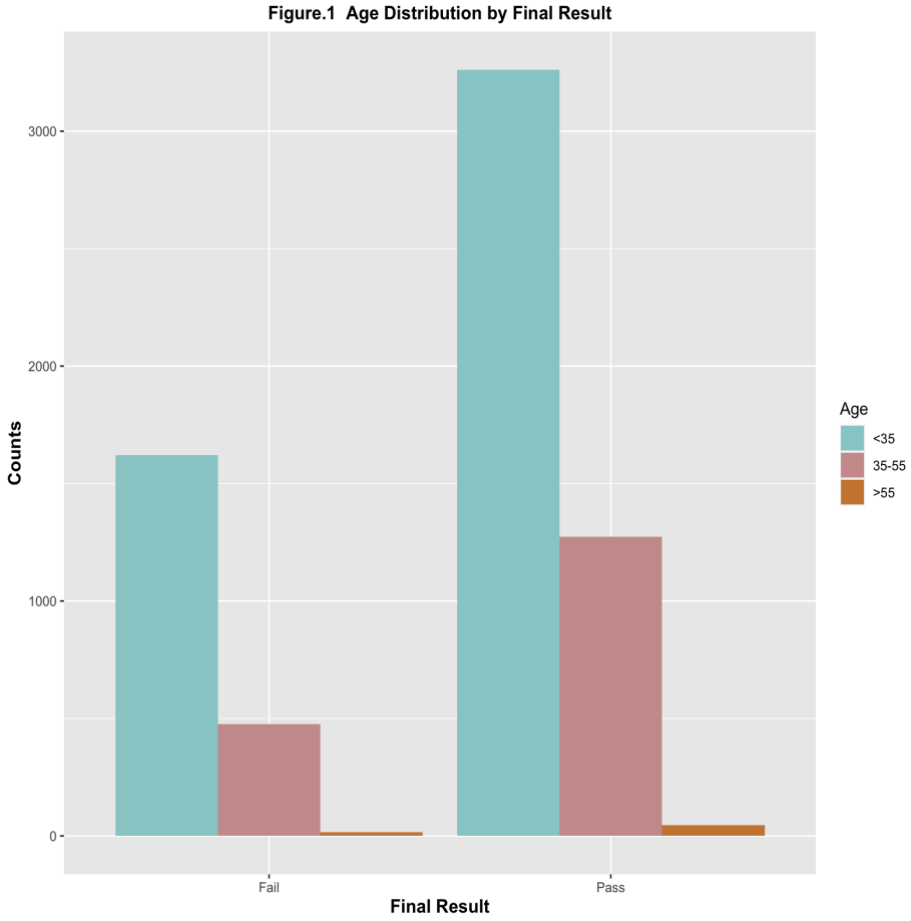
3. Exploratory Data Analysis （EDA）

Exploratory data analysis (EDA) will be executed on this STEM course dataset. This exploratory process analyzes both major categorical variables and numerical variables, which summarizes the distribution of the students' demographic information and students' engagement information on the courses. The major distribution results will be displayed using visualization methods, including bar plot, boxplot, and violin plot, which could create a more straightforward view of

datasets. EDA is essential for analysts to understand relationships among the key variables, which is beneficial to select features when building the predictive models in the next step.

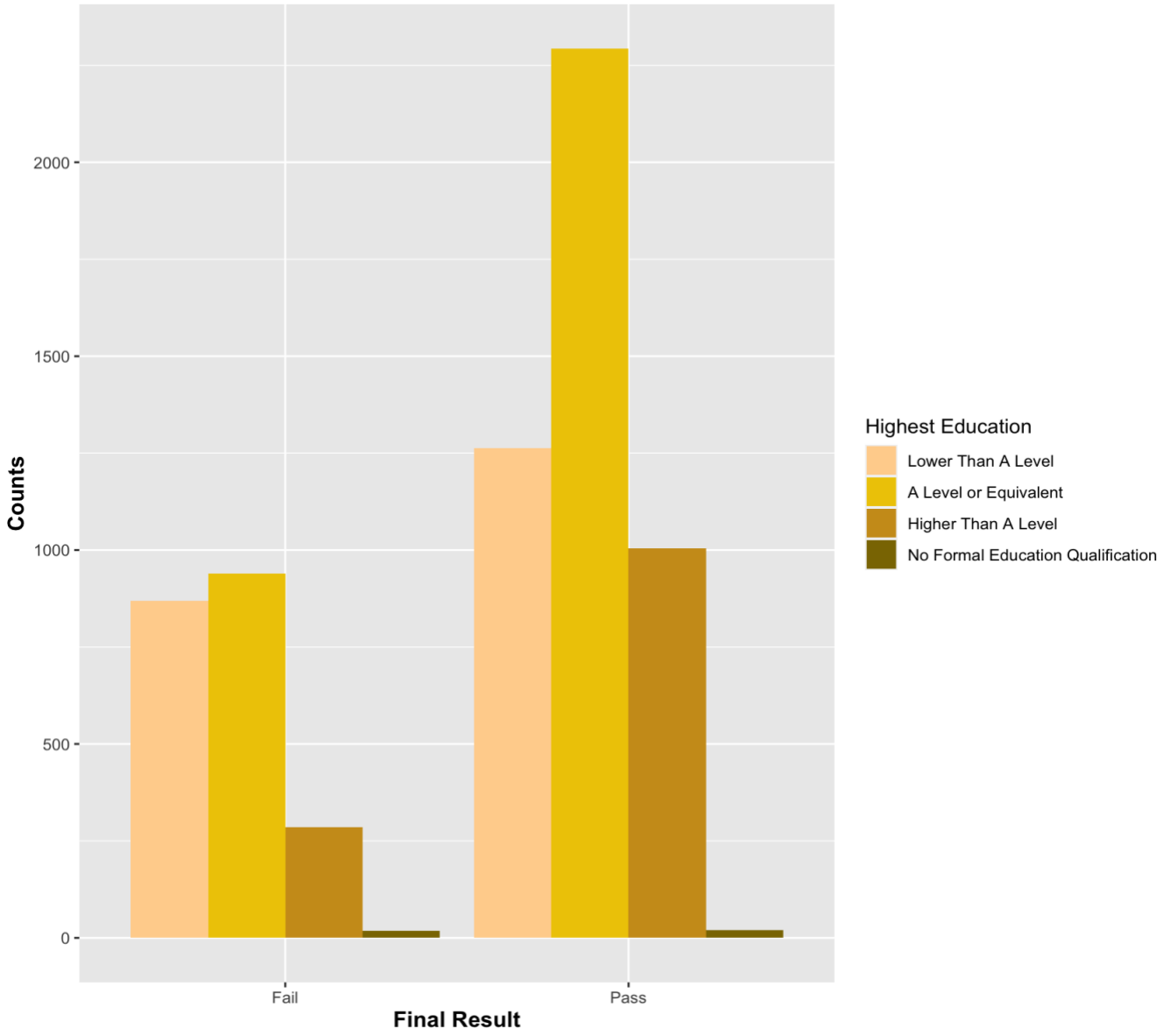
3.1 Student Demographic Information Analysis - Categorical Variables

Through exploring relationships between categorical independent variables of students' demographic information and the dependent variable of the final result, the trends and patterns between students' demographic data and the final result can be found. The primary relevant relationships are displayed through the following figures.



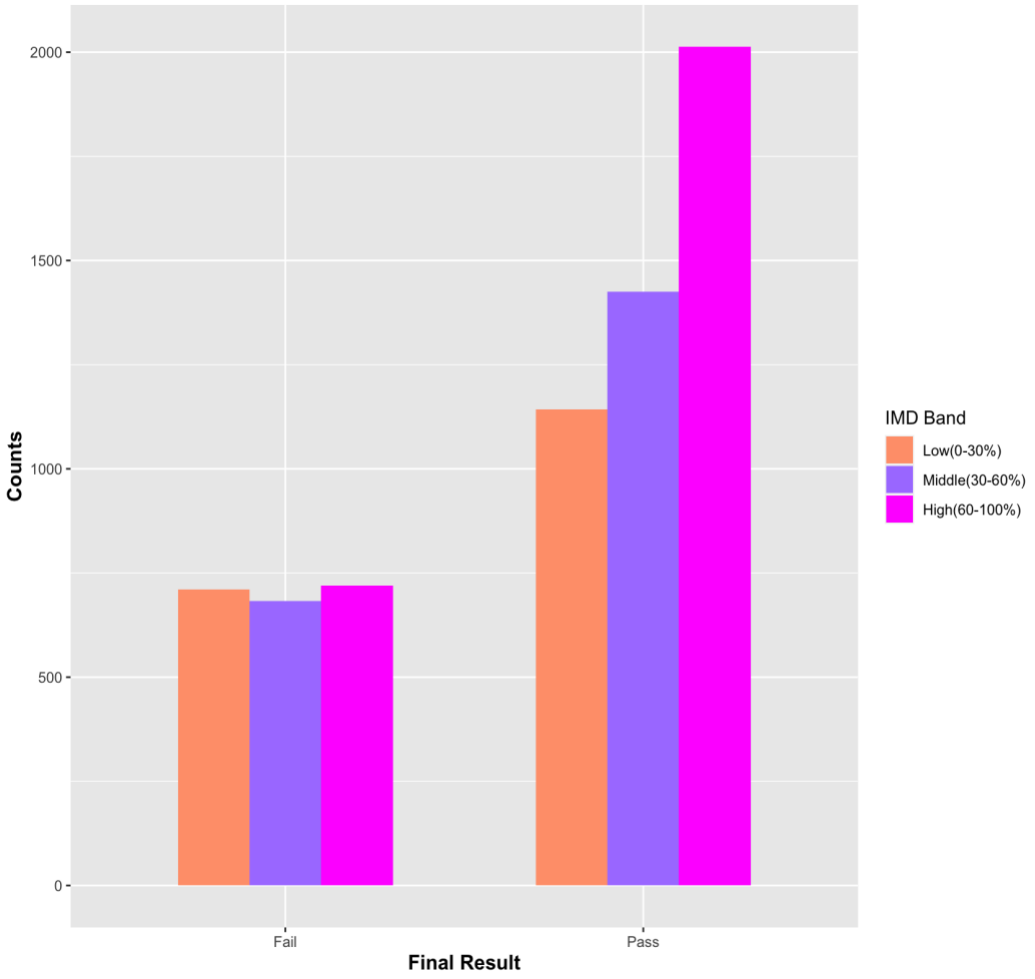
Through the figure 1, it can be found that students within the age group (<35) perform better in passing exams than those in age older than 35. It can be assumed that age might be one of the factors affecting students' performance on the STEM course.

Figure.2 Highest Education Distribution by Final Result



Through the figure 2, it can be found that students with high education levels (equivalent to or higher than A level) perform better in passing exams than those with education lower than A level. It can be hypothesized that one's educational level might be another factor influencing students' performance on the STEM course.

Figure.3 IMD Band Distribution by Final Result



Through the figure 3, it can be found that students with a high Index of Multiple Deprivation (IMD) perform better in passing exams than those with a low IMD index. Here, the IMD index is determined by a combination of factors, including “income, employment, health and housing, etc.”. The higher IMD, the better living conditions for humans.

3.2 Student Engagement Information Analysis - Numerical Variables

In addition to students’ necessary demographic data, students’ engagement data that indicates the degree of interacting with online courses is another primary consideration when predicting the final performance of students. The object is to explore if students’ click times, which represent the counts of students clicking a particular online course material, can affect their final results. Here, to better normalize the click times data distribution, log10() transform is applied for the y value.

Figure.4 Log of Clicktimes between Pass and Fail

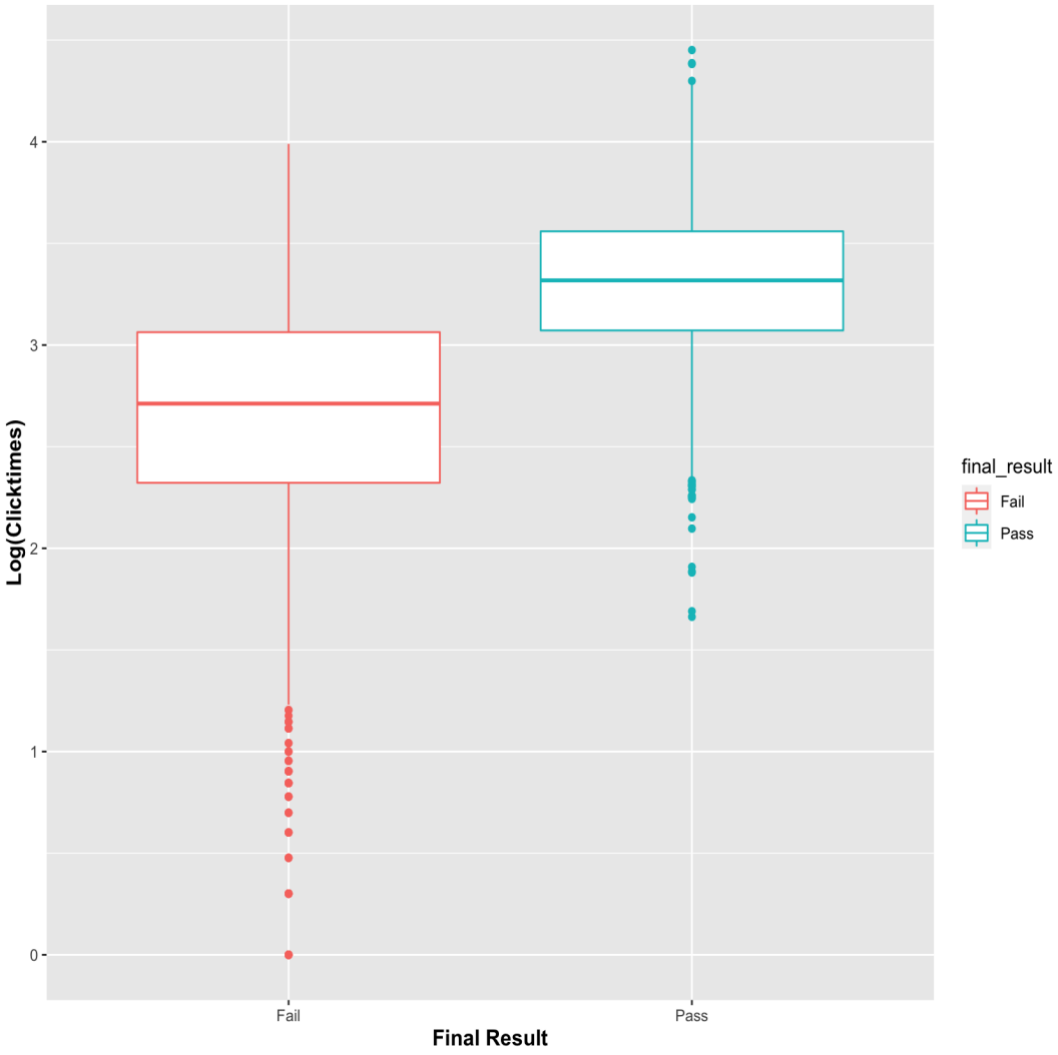
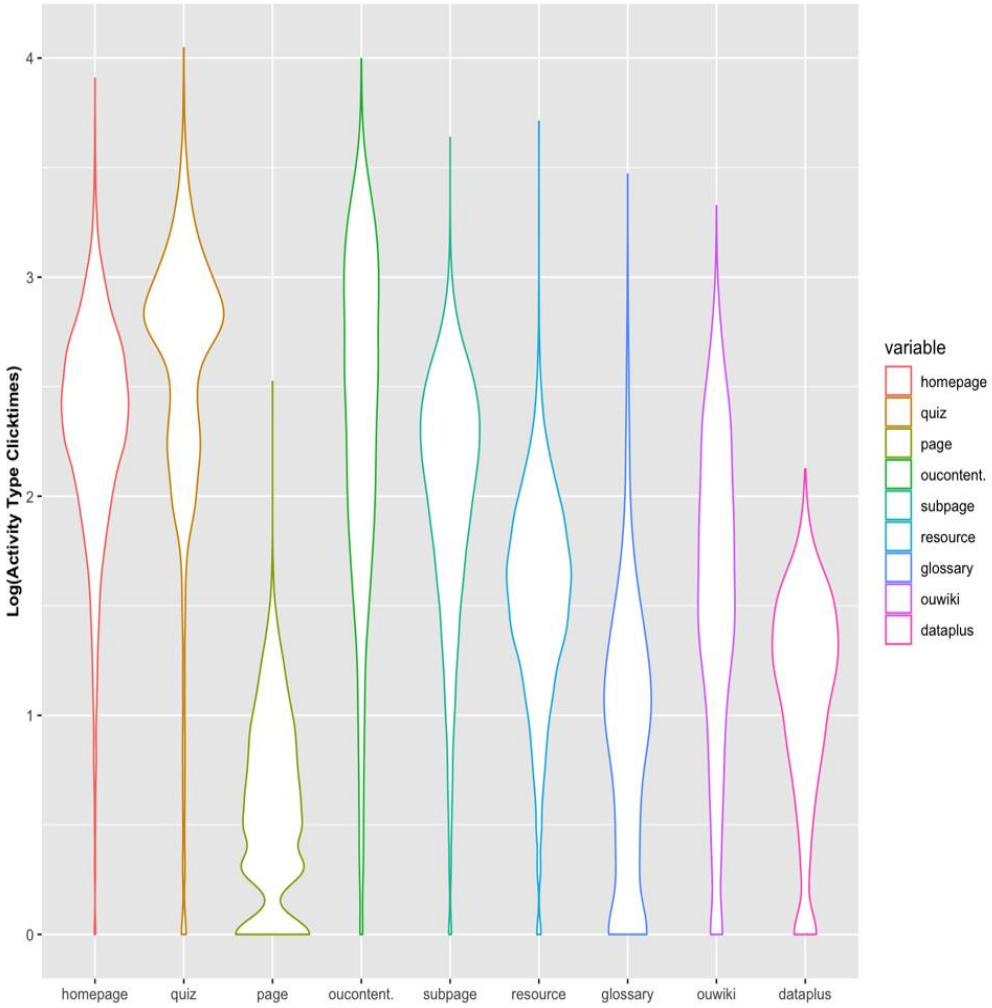


Figure 4 explores the relationship between the click times and the final result, which indicates that on the STEM online course, there is a significant difference in the click times between pass and fail results. Students who passed exams click far more times on the online course materials than ones who failed exams.

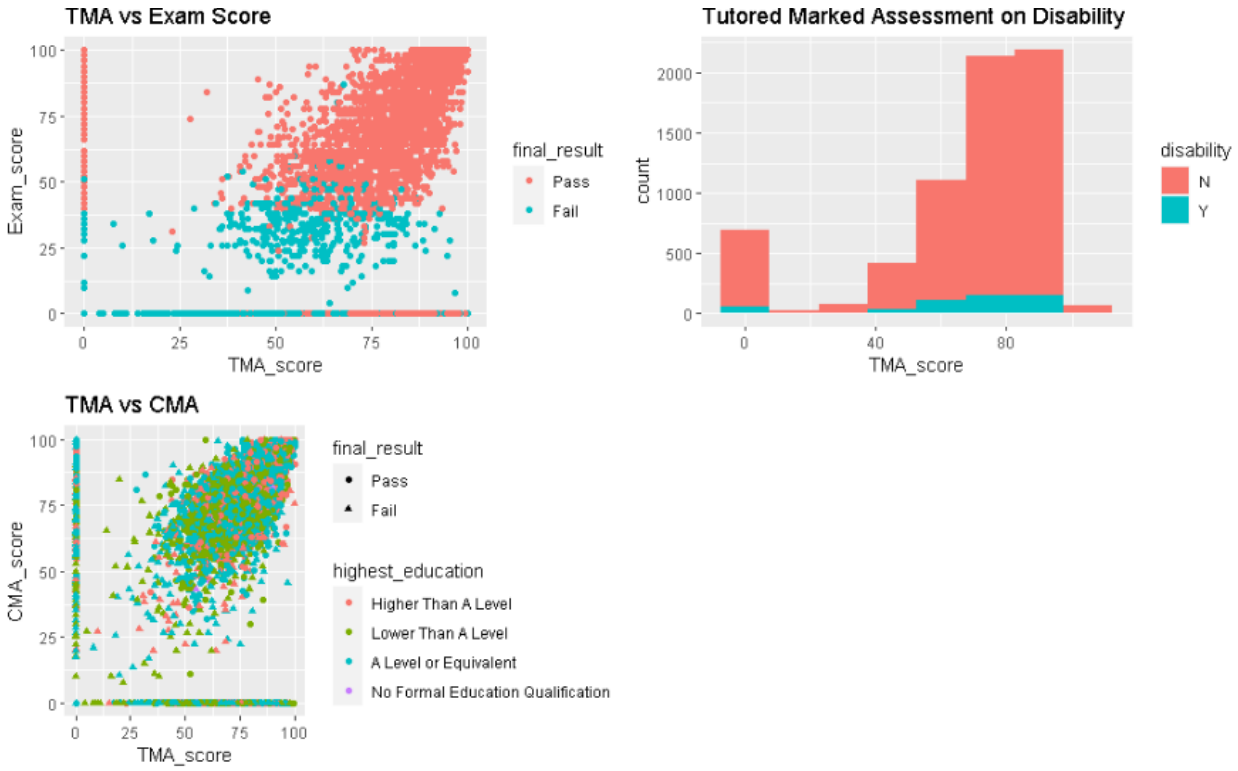
Then, different course characteristics determine that the STEM online courses are designed with various activity types online, which means that students can learn different courses by clicking corresponding web types. Following violin plot, *Figure 5*, is used here to display the click-time distribution of students interacting with different website activity types in STEM courses.

Figure.5 STEM Activity Types



From above Figure 5, it can be seen that STEM online courses are primarily presented by the web activity types of homepage, quiz, OUcontent (the open university course content website), subpage (a lower level web site on the open university website) as well as resource. During the learning process, students click more times on these web types to learn the online materials. As such, these web types are popular with being used by STEM coursed. Instead, among remaining activity types, fewer students learn online courses by clicking the "page" and "dataplus (a web-based tool for enabling students an efficient method of reviewing classes)" types. Understandably, different web type designations have different attraction levels for students to click.

Besides students' engagement activities – such as clicking online courses mentioned above affect their final results, scores and evaluation assessed by tutors and computers are also affecting the students' final results to a great extent. Following figures explore distributions about students' scores data.



The above plot is made of 3 smaller plots. The top left plot shows that students who passed tended to have higher TMA and Exam scores. On the other side, the top right plot indicates that majority of students scored between 70 and 100. Of those students, very little were those with handicaps. Lastly, the bottom left plot shows the relationship of the TMA (Tutored Marked Assessment) to CMA (Computer Marked Assessment). The cluster of points are students that passed and have quality of formal education.

4. Unsupervised Analysis

4.1 Cluster Analysis

As another technique employed in the exploratory data analysis, clustering is an unsupervised machine learning algorithm for grouping similar observations into a number of clusters based on multiple variables for each individual observed value. The purpose of cluster analysis or clustering is to group a collection of objects in such a way that objects in the same group (called a cluster) are more similar to each other (in some sense) than objects in other groups (clusters).

In this project, two clustering algorithms, K-means and hierarchical, will be applied and evaluated first; the optimal between these two algorithms will then be selected for the cluster analysis on this dataset.

Before carrying out a cluster analysis, it's important to concern about data standardization. Available cluster analysis algorithms all depend on the concept of measuring the distance (or some other measure of similarity) between the different observations that will be clustered. If one of the variables is measured on a much larger scale than the other variables, then whatever measure that is used will be overly influenced by that variable. In this dataset, numerical variables are measured on different scales. For example, *Exam_score* is measured on the 0-100 scale, while some value of the *Sum_click* is far more than 100. Thus, it's essential to scale data before cluster analysis. Then, this study focuses on analyzing two final results, pass and fail, of the students. As a result, the initial number of clusters, *k*, will be identified 2. The K-means and hierarchical clustering algorithms will be evaluated with the case of k=2 on the scaled dataset.

4.2 Cluster External Validation

K-means		
Cluster number	**Cluster 1**	**Cluster 2**
Final result - Fail	86	2026
Final result - Pass	1263	3318

Hierarchical (Complete Linkage)		
Cluster number	**Cluster 1**	**Cluster 2**
Final result - Fail	2112	0
Final result - Pass	4580	1

Through evaluating external validation, it can be found that the Hierarchical-clustering result is much better than that of K-means. Then, internal validation would be made next.

4.3 Cluster Internal Validation

Cluster Internal Validation (K=2, Scaled)		
Internal Metrics	**K-means**	**Hierarchical (*Complete Linkage*)**
tot. withinss	114086.8	129771
average. between	7.254587	63.94126
average. within	5.2995	5.719017
ave. silwidth	0.2650943	0.9104142
Dunn Index	0.007364709	0.5785261

Through evaluating internal validation by different metrics between K-means and hierarchical, it can be found that Hierarchical is the better model as it does well on 3 categories (average.between, avg.silwidth and Dunn index), rather than 2 categories that the K means fares well in (total. Within SS and Average Within SS)

4.4 Cluster Selection and Improvement

Through above both external and internal cluster validation, it can be decided that hierarchical clustering using complete linkage is the optimal cluster algorithm at this dataset. As a result, hierarchical clustering algorithm will be used to carry out cluster analysis in the following. To further improve the performance of hierarchical clustering, the method of feature selection and cluster number adjustment will be tried on the hierarchical clustering.

Internal Metrics	Single	Complete	Average	Centroids	Wards
tot. withinss	131317.2	129771	129771	129771	117803.5
average. between	50.42	63.94	63.94	63.94	7.12
average. within	5.72306	5.719017	5.719017	5.719017	5.393955
ave. silwidth	0.89	0.91	0.91	0.91	0.24
Dunn Index	0.570	0.579	0.579	0.579	0.011

To determine which algorithm of hierarchical clustering was the best, multiple methods were tested. The above table displays the results from the various algorithms used. After running each of the algorithms the results were compared to see which method had the best score in each category. Complete, Average, and Centroids all produced the same results with this data set. Since they were all the same, Complete linkage was chosen as the method to represent this group. These methods had the best metrics for average between, dunn index, and average silhouette width. Ward's method had the best metrics for total within sum of squares and average within. With complete linkage having the best result in three of the five metrics, over ward's winning in two, complete linkage was chosen as the best method.

Also, as final result is the dependent variable in this dataset, it was used to determine the optimal number of clusters. Since, final result is a two-category variable, *k* was set to 2. This allowed the

model to split into two clusters and use those clusters to compare with the dependent variable for accuracy.

4.5 Cluster Conclusion

Comparing the two methods between K-means and hierarchical clustering led to the following interesting results. For both models, k = 2 was used due to the dependent variable having two categories. Based on the 5 metrics tested for cluster validation, hierarchical scores slightly better. Hierarchical has a better average between, average silhouette, and Dunn index scores. K-means has a better total within sum of squares and average within. Since hierarchical scored better on three of the metrics, that was chosen as the optimal clustering algorithm in this project.

5. Supervised Analysis

5.1 Model Introduction

For this section, supervised learning will be used. In the previous sections of exploratory analysis and cluster analysis, the analysis of relationships of features like the number of clicks and *imd_band* with the variable *final_result* was done. The *final_result* is the variable that represents the student's passing or failure of the course they were enrolled in. In this project, the goal is to build a model that could perform best and have the best goodness of fit when predicting a student's final result and classify students into two groups well. Therefore, model performance metrics, such as accuracy, F1, recall, etc. will be the primary factor determining which one model is the most appropriate.

For this section, the following models will be compared: Decision Tree, Random Forest, Naive Bayes, Artificial Neural Network, and K-Nearest Neighbors (k-NN) as they are candidate models for predicting the dependent variable, *final_result*.

5.2 Model Comparison

To determine the optimal model, the metrics of performance and goodness of fit are evaluated on each model with respect to training and testing (80-20), as the following figure shown. The general evaluation metrics include Accuracy, Sensitivity (True Positive Rate), Specificity (True Negative Rate), Kappa, and F1 score. However, as models tend to score higher on some of the categories, a weighted average metric known as "Score" was computed. This metric will be the sole determinant on ranking the models. Each metric was given a weight of 0.2.

$$Score = 0.2\,(Accuracy + Sensitvity + Specificity + Kappa + F1)$$

Model	Accuracy	Sensitivity	Specificity	Kappa	F1	Score
Decision Tree	91.26%	96.94%	78.91%	78.93%	93.82%	87.97%
ANN	90.28%	95.31%	79.38%	76.85%	93.07%	86.98%
Naive Bayes	75.71%	69.21%	89.81%	51.20%	79.60%	73.11%
Random Forest (Win!)	**92.97%**	**96.18%**	**86.02%**	**83.48%**	**94.94%**	**90.72%**
K-NN	78.55%	93.23%	46.68%	44.55%	85.61%	69.72%

5.3 Model Selection

From the table in the previous section, one can see that Random Forest yields the highest score. Therefore, the Random Forest model will be selected and explored further in the following supervised learning analysis. The following table pertaining to the performance of the Random Forest model on the original dataset.

Model	Accuracy	Sensitivity	Specificity	Kappa	F1	Score
RF (Train)	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
RF (Test)	92.97%	96.18%	86.02%	83.48%	94.94%	90.72%

It can be found that, when the Random Forest is explored deeper through looking at the training and testing metrics respectively, the model has overfitted. Improving and optimizing the model is necessary to hopefully prevent this problem.

5. 4 Model Improvement

To improve the Random Forest, the following methods are implemented:

- RFE (Recursive Feature Elimination)
- Sampling to balance the training set
- Hyperparameter tuning

5.4.1 RFE (Feature Selection)

To start, the recursive feature elimination, 10-fold cross-validation, repeated 3 times is induced on the set. 16 variables are chosen to be the optimal amount, but it is still considerably large. It is possible that the model won't be able to predict well given this amount, hence it is necessary to select the top 5 out of the 16 variables. As a result, 5 variables, including *Exam_Score*, *TMA_Score*, *Ouwiki*, *Sum_clicks* and *Quiz* are chosen.

5.4.2 Balanced Sampling

An imbalance dataset may be the reason the model performed poorly on the testing set. It is vital to balance the training set so there is an equal number of "pass" and "fail" observations. Under-sampling the majority class ("pass") rather than oversampling the minority class ("fail") will be chosen because the training set is larger than 1000 observations and overfitting should be minimized as much as possible.

5.4.3Hyperparameter Tuning

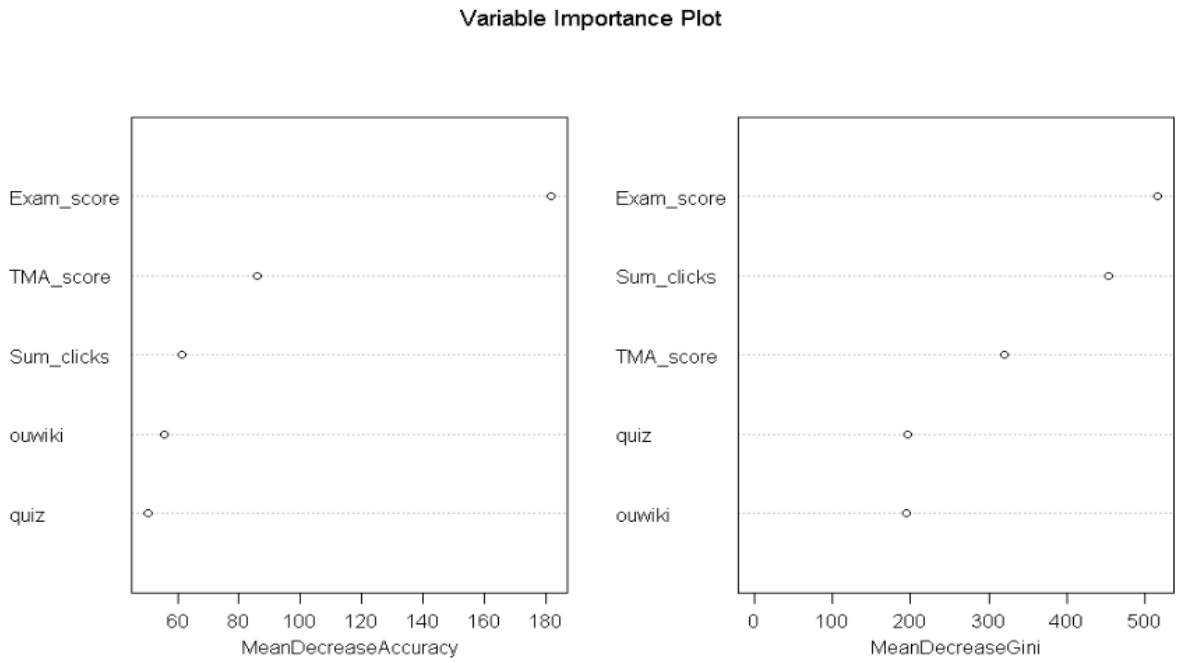
Both Grid search and Hyperparameter tuning were evaluated to tune the Random Forest model. To analyze the Random Forest model, instead of using the grid hypertuned model, which isn't actually a random forest model, a Random Forest model that was trained on the featured set of 6 variables mentioned earlier was used. Recall that the Random Forest was chosen out of the 5 models because of its metrics and better fit. In addition, consideration of choosing the model stemmed from its less complex natured as compared to a model like ANN. The Random Forest is made of multiple decision trees, and each of these trees gives a decision on what classification a certain observation should be assigned based on its own set of decision rules. Ultimately, the final decision is made due to the majority of votes. If 80% of the trees decide that an observation should be assigned a certain label, then that observation will be classified with the label. Because the forest uses multiple trees, this explains why it was able to defeat the Decision Tree model in specificity and sensitivity.

There is also a flaw in this model. Because the Forest uses multiple trees, one can't visualize it or the rules. Denial of model logic is exchanged for better classification metrics.

5.5 Model Discussion

The Random Forest model finally determined the following three variables as most important, as the following plots shown:

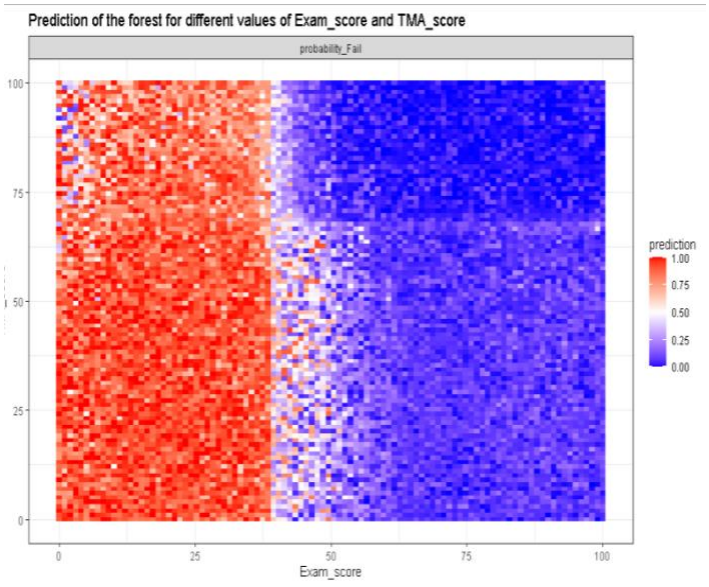
- TMA Score
- Exam Score
- Sum_clicks

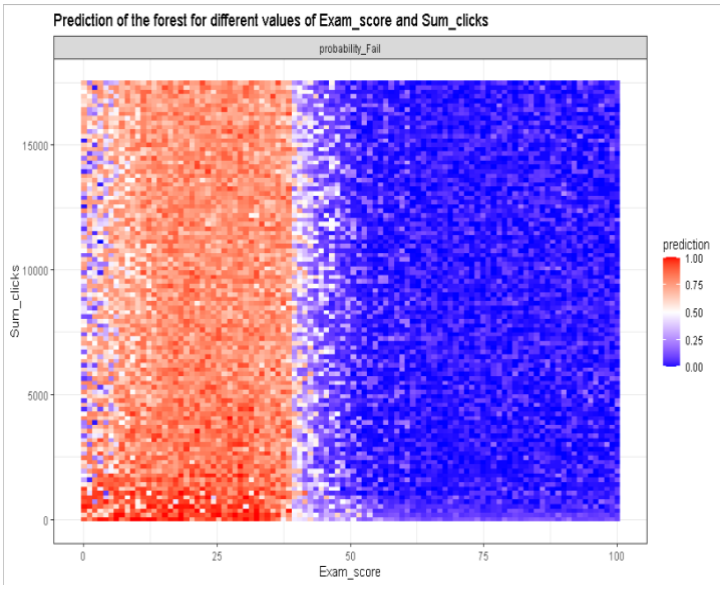


From above plots, it can be seen that the variance important plot contains Mean Decrease Accuracy and Mean Decrease Gini. These components are used to evaluate the importance of the variables.

- MeanDecreaseAccuracy: It represents the amount of decrease in accuracy if a feature happens to be removed from the model. In the diagram below, Exam_Score seems to have a high level of importance and would cause a huge reduction in classification accuracy should it be removed. The top 3 variables in this category are Exam_Score, TMA_Score and Sum_Clicks.
- MeanDecreaseGini: It refers to the average decrease in node impurity. A higher decrease in node impurity means that the observation has a higher chance of been classified correctly by the model in this case Random Forest. The MeanDecreaseGini ultimately lets us know which variables are important in causing the greatest decreases of gini impurity, the importance of the variables in predicting the dependent variable. It can be observed that the same three variables cause the most in gini impurity decrease: Exam_Score, TMA_Score and Sum_Clicks

Then, it's essential to examine the relationships of the top 3 features having a significant impact on the dependent variable. The diagram below is similar to a heatmap. It shows the probability of each observation with relation to that observation having a status of "fail".





It can be seen that in both plots, Exam_Score is a determining factor between passing or failing a course. Students that have an Exam score of 0.75 and 1 have a higher probability of passing the course as opposed to exam scores of 0 and 0.4. Any score between 0.4 and 0.5, the model deems a 50% probability in failing.

This makes sense as students who score high in exams tend to pass the course and those with poor scores tend to flunk. Of course, there is the occasional on the fence student who can pass if he or she works hard enough or fails if they decide to give up. The outcome is uncertain for this type of student. However, in both diagrams, when Exam_Score =0, some observations are considered as "passing" according to the model. One may attribute this to the model's error rate in incorrectly classifying the observations.

5.6 Model Findings

From the supervised analysis, it was derived that the number of clicks, the score on the final exam, and the score on the Tutored Marked Assessment were key indicators of determining if a student passed or failed their course. It would make sense why the Random Forest model chose these variables. The sum of clicks represents the student's consistent engagement throughout the course. Student engagement or participation combined with good exam scores, especially those before the termination of a course, are usually the important factors that cause a student to pass or fail.

6. Summary

6.1 Discussion

In addition to the important variables (such as Sum_click and TMA_score which affect final grades) that the Random Model finds out, perhaps there are other external factors affecting students' final performance. For example, in the exploratory data analysis section, students were found for the most part to have come from medium or high-income families. Income might be a deciding factor whether a student passes or fails. As Meredith Kolodner's article mentioned that low-income college students already faced barriers to graduating and they have a tough time passing courses in an online learning environment. She writes the following:

> "For low-income students, though, the situation can be dire. Earning a degree is challenging in the best of circumstances — graduation rates for low-income students have remained stubbornly low for decades. Only 14 percent of the lowest-income students earn a bachelor's degree within eight years of first enrolling, according to the most recent government data. Juggling bills, jobs and family responsibilities can make it difficult to find the time and headspace to study. The coronavirus pandemic is exacerbating those pressures. Students who lived on campus are trying to keep up their

> grades at home, some in cramped or emotionally tense living conditions. Adult students who have children are being buried by home-schooling demands." (Kolonder, 2020)

Therefore, it's essential to combine the factor of the students' individual hard-working extent with the outside environment factors to evaluate and assess their study performance.

6.3 Conclusion

To summarize, this project presents a contribution to knowledge in early prediction of students' performance in online courses, determining students likely to fail or pass the STEM online courses. Results reveal that some student's demographic characteristics and student's engagement activities, including clicking online courses and participating exams, have a significant impact on student performance. Students within the age group (0-35), in a high education level, and in a better living condition performed better in passing exams. Also, the relationship between click times and the final result indicates that on STEM online course, students who passed exams click far more times on the online course materials than ones who failed exams.

This study also determines the effectiveness of the unsupervised model and supervised model in the early prediction of student performance, enabling timely intervention by the professor to implement corrective strategies for students' support and counseling. For the unsupervised model, clustering with complete linkage algorithm performs best in grouping students. For the supervised model, the random forest model outperforms the other supervised models in predicting students' performance in STEM domain online courses.

In short, a good education is necessary for everyone to improve knowledge, economic status, as well as lifestyles. Currently, online learning is a convenient way to access educational resources so that people can acquire knowledge at anytime and anywhere.

Appendix

Appendix A - Variables description

Variables	Names	Types	Description
V1	*id_student*	Numerical	The unique id number of each student
V2	*code_presentation*	Categorical	Code name of the presentation. It consists of the "2014B-STEM" for the presentation of STEM course starting in February and "2014J-STEM" for the presentation starting in October.
V3	*gender*	Categorical	Student's gender
V4	*highest_education*	Categorical	The highest student education level on entry to the module presentation.
V5	*Imd_band*	Categorical	The IMD band of the place where the student lived during the module-presentation. Here, it consists of three levels of IMD index for the presentation of different conditions of income, employment, health, living environment. The higher, the better.
V6	*age_band*	Categorical	Age levels
V7	*disability*	Categorical	Indicates whether the student has declared a disability. (Y is yes, N is no)
V8	*date_registration*	Numerical	The day of student's registration for the module presentation.

V9	*Exam_score*	Numerical	Final exam score (an invigilated three-hour paper).
V10	*TMA_socre*	Numerical	Tutor marked assessment score (writing problem).
V11	*CMA-score*	Numerical	Computer marked assessment score (six multiple-choice or "missing word" 10-question).
V12-V20	*Name of activity type*	Numerical	The total number of times the student interacted with different kinds of online materials.
V21	*Sum_clicks*	Numerical	The total number of times a student clicks online materials in the entire course period.
V22	*final_result*	Categorical	Student's final result in the module-presentation. There are two kinds of results: Pass (score>=70) and Fail (score < 40).

References

1. Kuzilek, J., Hlosta, M. & Zdrahal, Z. Open University Learning Analytics dataset. *Sci Data* 4, 170171 (2017). https://analyse.kmi.open.ac.uk/open_dataset

2. Kolodner, Meredith, and The Hechinger Report. "Low-Income College Students Already Faced Barriers to Graduating. The Coronavirus Multiplied Them." *The Washington Post*, WP Company, 8 Apr. 2020, www.washingtonpost.com/education/2020/04/08/low-income-college-students-already-faced-barriers-graduating-coronavirus-multipied-them/

3. Peach, R. L., Yaliraki, S. N., Lefevre, D., & Barahona, M. (2019). Data-driven unsupervised clustering of online learner behaviour. *Npj Science of Learning*, *4*(1). doi: 10.1038/s41539-019-0054-0
4. Marbouti, F., Diefes-Dux, H. A., & Madhavan, K. (2016). Models for early prediction of at-risk students in a course using standards-based grading. *Computers & Education*, *103*, 1–15. doi: 10.1016/j.compedu.2016.09.005
5. Minaei-Bidgoli, B., Kashy, D., Kortemeyer, G., & Punch, W. (n.d.). Predicting student performance: an application of data mining methods with an educational web-based system. *33rd Annual Frontiers in Education, 2003. FIE 2003.* doi: 10.1109/fie.2003.1263284
6. Hu, Y.-H., Lo, C.-L., & Shih, S.-P. (2014). Developing early warning systems to predict students' online learning performance. *Computers in Human Behavior*, *36*, 469–478. doi: 10.1016/j.chb.2014.04.002
7. Everitt, B.S., Landau, S. and Leese, M. (2001), Cluster Analysis, Fourth edition, Arnold.