MGT 6203 Group Project Final Report – Team #8

Hugo Dupouy (hdupouy3) - Brendan Danyluik (bdanyluik3) - Jacobo De Leon (jgarcia340) - Henning Garvert (hgarvert3) - Michael Yang (myang387)

PROJECT OVERVIEW

Project Title: Predicting student performance based on gameplay

Background Information on chosen project topic

The "Predict Student Performance from Game Play" project is a Kaggle-hosted educational data analysis challenge concerned with forecasting if a student will correctly answer a question based on past performance in the Jo Wilder-created online educational game (Franklin, et al., 2023). We believe that gamification will play an essential role in the future education of children (Wang, Chen, Hwang, Guan, & Wang, 2022) but is also of great relevance in the context of professional education. Indeed, it has been shown that motivational and knowledge enhancement increased significantly while integrating an educational game into classrooms (Manzano-León, et al., 2021; Nietfeld, 2020).

Problem Statement

This project aims to enhance student learning outcomes by assisting educators and game developers in better understanding the connection between gameplay and academic performance. The dataset for the challenge contains data on student gameplay statistics, such as the number of levels cleared and time spent playing it.

Hypothesis

Our hypothesis is that it is possible to predict whether a student will pass a particular question based on external factors. In the context of the game, we expect that variables such as 'elapsed time', 'level of the game', 'hover duration' or whether the game's music has been turned on or off do impact the game results. We aim to figure out what degree of variability in the data can be explained by our models.

Research Questions

1. Can future student performance be predicted from past student performances in an online educational game?
2. Can external factors have an impact on student performances in an online educational game?

Business Justification

By evaluating gaming data metrics, participants in the challenge can learn essential lessons about what influences student achievement in the Jo Wilder game and possibly in other educational environments. The initiative is a crucial step in enhancing learning outcomes for students worldwide by creating more valuable and exciting educational games and tools and providing opportunities for improving other existing educational games and tools. Indeed, Manzano-León et al. (2021) highlight that gamification would increase children's motivation toward education. Also, Nietfeld (2020) mentioned that educational games would improve children's knowledge, especially for highly motivated children, because these players who gained the most significant knowledge from the game were also the most willing to use it in other contexts.

Besides the learning aspect of the challenge, the ability to predict someone's performance based on past results as well as certain boundary factors can additionally be conveyed to a business environment, where people constantly need to make decisions. In such cases, a proper model can help in two ways: Identifying people and specific situations, where an increased risk of making wrong decisions is given or, identifying external factors, which are increasing the probability of making wrong decisions. This can be especially relevant if it is not possible to track or access how a specific person has performed in the past.

DATA PROCESSING

Data Sources

https://www.kaggle.com/competitions/predict-student-performance-from-game-play/data

Data Description

Data on gaming metrics for a group of students who played the Jo Wilder educational game are included in the "Predict Student Performance from Game Play" dataset in Kaggle. The dataset contains 26,300,674 records and 20 columns (training and testing set together to take up about 4.70 GB). In general, gaming data is provided, such as the number of levels finished and the time spent playing. We attached a screenshot of the "Test" table in Appendix A.

For data cleaning, the data had been pre-cleaned by the authors and/or Kaggle's administration. Data points that were not formatted incorrectly, incomplete, or entered wrong. As some players only did parts of the game, we filtered out the respective session_ids. Additionally, we needed to transform some columns to factors such as session_id, level, and coordinates. To process the data, we have used the file type ".Rmd" to simply run code and extract a well-formatted html file that would be easier to read in the submission process the competition.

While the data did not need to be cleaned for incorrect formatting or typos, the data did require handling of NA values, as the nature of the features relating to events or clicks within certain portions of the game (Whether that be in the journal, or only occurring when hovering on items) created large volumes of NA values in key features such as hover_duration, page, room_coor_x, room_coor_y, screen_coor_x, and screen_coor_y. Handling these NA values was tricky, as a standard values of 0 or -1 were meaningful values for these parameters Instead, very large or unique values were used to ensure that the unique game states were captured. This allowed for handling of NA values while not skewing pre-existing counts of values.

Key Variables

The independent variables include gaming metrics like the number of levels finished, the amount of playtime, current level, click coordinates, the event's name, the text seen during the event, whether music is on, and if the game is in high-quality mode (Appendix B). In addition, we speculated that some factors might be more significant than others in forecasting student performance based on prior knowledge and intuition. For instance, considering that they show the student's exposure to and involvement with the game, we anticipated that the number of levels finished and the amount of time spent playing would be excellent performance indicators.

Explanatory Data Analysis

We performed explanatory data analysis to better understand the key features of the data, identify potentially required data transformations, and check the meaningfulness of the parameters.

At first, we created a correlation matrix of all the variables by changing the character variables into factors, then into numerics, and filtered out any variables with perfect correlations (Figure 1). We also checked the significance of these variables with a correlation factor greater than 0.5 to determine a complete and clean correlation matrix that kept only six variables (index, event_name, name, text, text_fqid, and level). This correlation matrix's main takeaways are the negative relationship between 'name' and 'text'/' text_fqid' with -0.62, indicating that the text displayed negatively correlates to the different event names. However, the event type's name is positively correlated to each text_fqid.

Additionally, we determined the distribution of the event names by counting their number of actions (Figure 2). It shows that the students tend to navigate a lot with clicks (~43%) and to enquire about in-game characters (~23%). However, students tend to disregard observation_click (~0.81%), indicating that they do not try to search a lot in the environment scene.

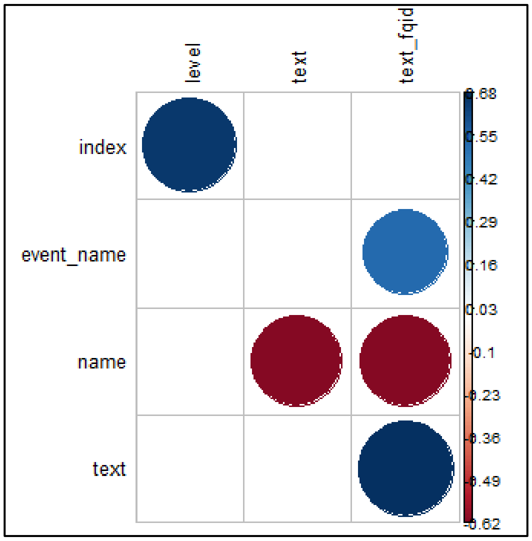


Figure 1 - Correlation Matrix

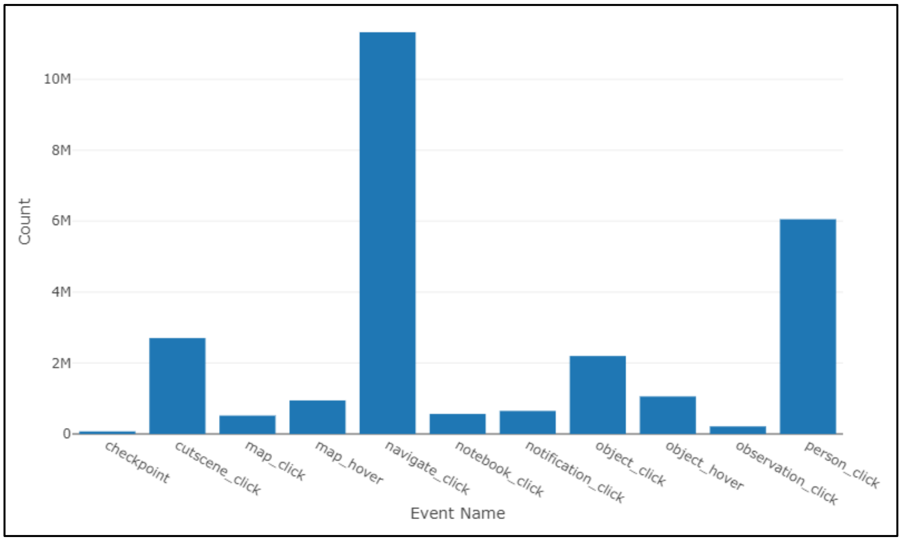


Figure 2 - Distribution of the Event Names

Furthermore, visualizing the accumulated mean time the players need to complete each level shows that this time tends to increase from level to level, and achieving higher levels tends to take more time than lower levels in the students' game progression (Figure 3). Moreover, it would indicate that students tend to complete the game in 22 hours and 47min on average.

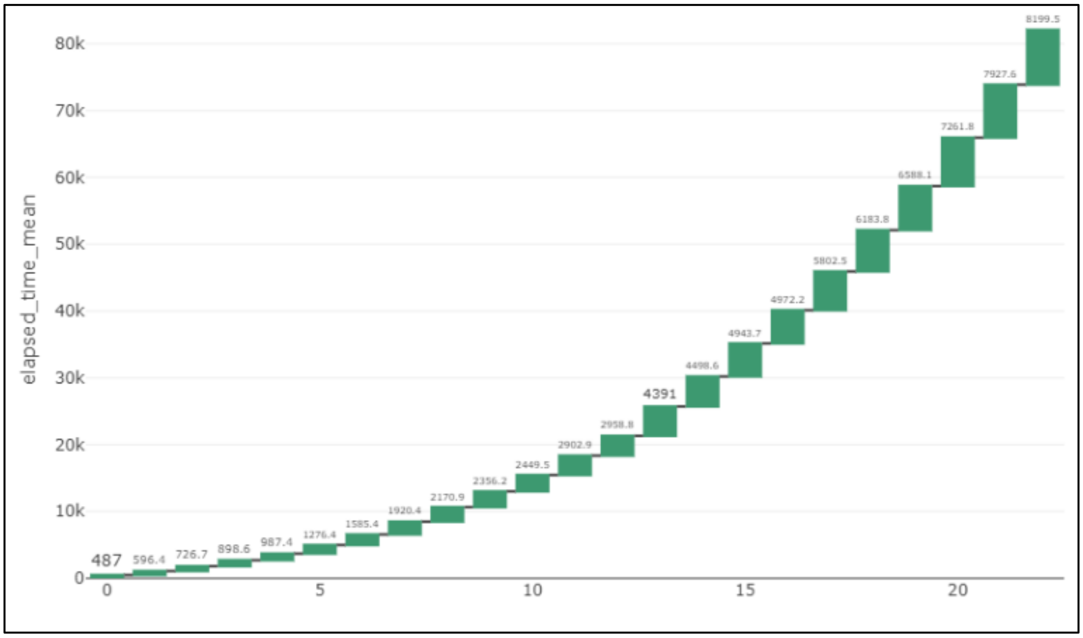


Figure 3. Cumulated Elapsed Time Mean in the level progression

METHODOLOGY

Applied Approach

The exploratory data analysis allowed us to look at how our data is distributed and if there are any correlated columns. After exploratory data analysis, Factor Analysis of Mixed Data (FAMD - available in the FactoMineR package) was used to look at combining or altering our independent variables. FAMD was used instead of PCA (Principal Component Analysis), as it is better suited to handle qualitative and quantitative data types (Visbal-Cadavid, Mendoza-Mendoza, & De La Hoz-Dominguez, 2020).

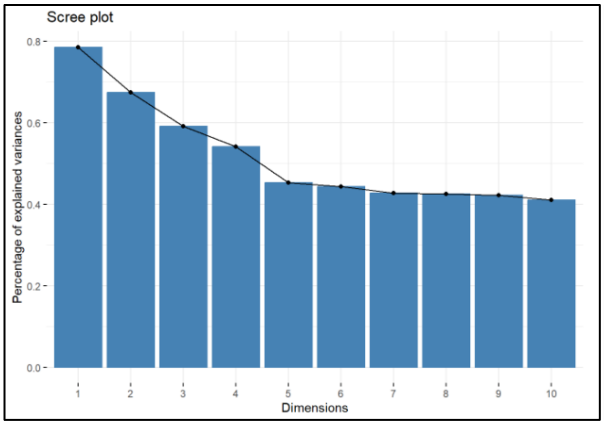


Figure 4. Scree Plot of FAMD

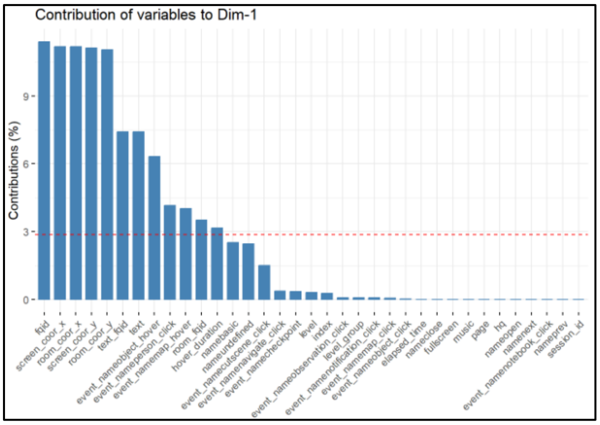


Figure 5. Result of FAMD Analysis for Dimension 1

Our data contains a mixture of continuous, discrete, and categorical data. However, we experienced that FAMD is more computationally intensive, and due to memory restrictions, the dataset for the FAMD was sampled with a total of 10,000 data points. Even after trying 100,000 data points, we found similar results. Therefore, to focus on the most relevant variables, we decided to identify those dimensions of variables that are suitable to explain the highest degree of overall variability. An overview of the results can be seen in Figure 4, which shows us how much of the variability can be explained by the different sets of dimensions. Indeed, the first dimension explains up to 80% of the variability. The dimensions ranked second and third are based on fewer variables, as seen in Figures 5-7. However, these dimensions explain less variability in the data. Therefore, we decided to focus on the highest-ranked dimension as it explains almost ~80% of the variance in the data.

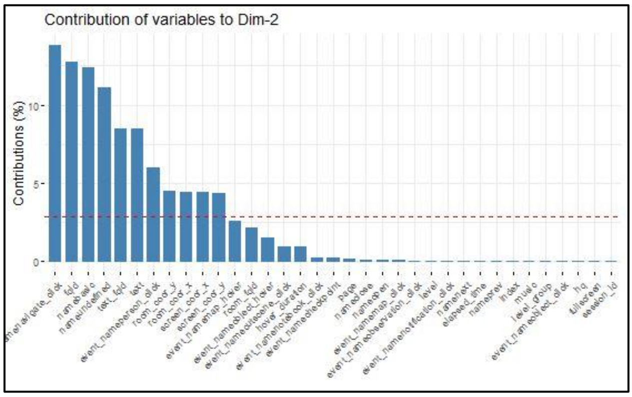


Figure 6. Result of FAMD Analysis for Dimension 2

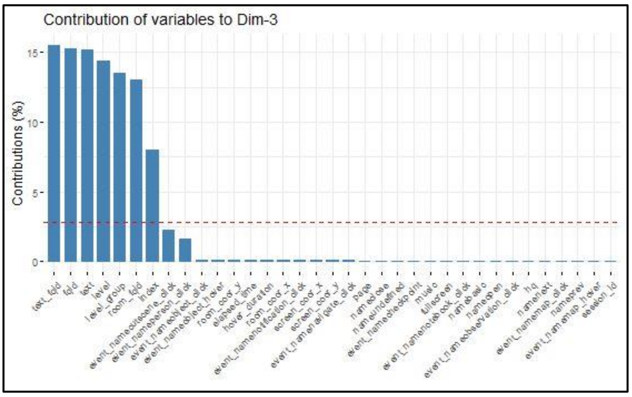


Figure 7. Result of FAMD Analysis for Dimension 3

The biplot and cos2 plot (Figure 8) visualizes that the variables are all positively correlated between dimension 1 and dimension 2, and some variables are better represented than others. Based on this principal component method, we determine the quality of representation of each variables in dimension 1 and 2. We made sure that the variables highlighted in green were kept when filtering for the variables in Dim$1 to run the models.

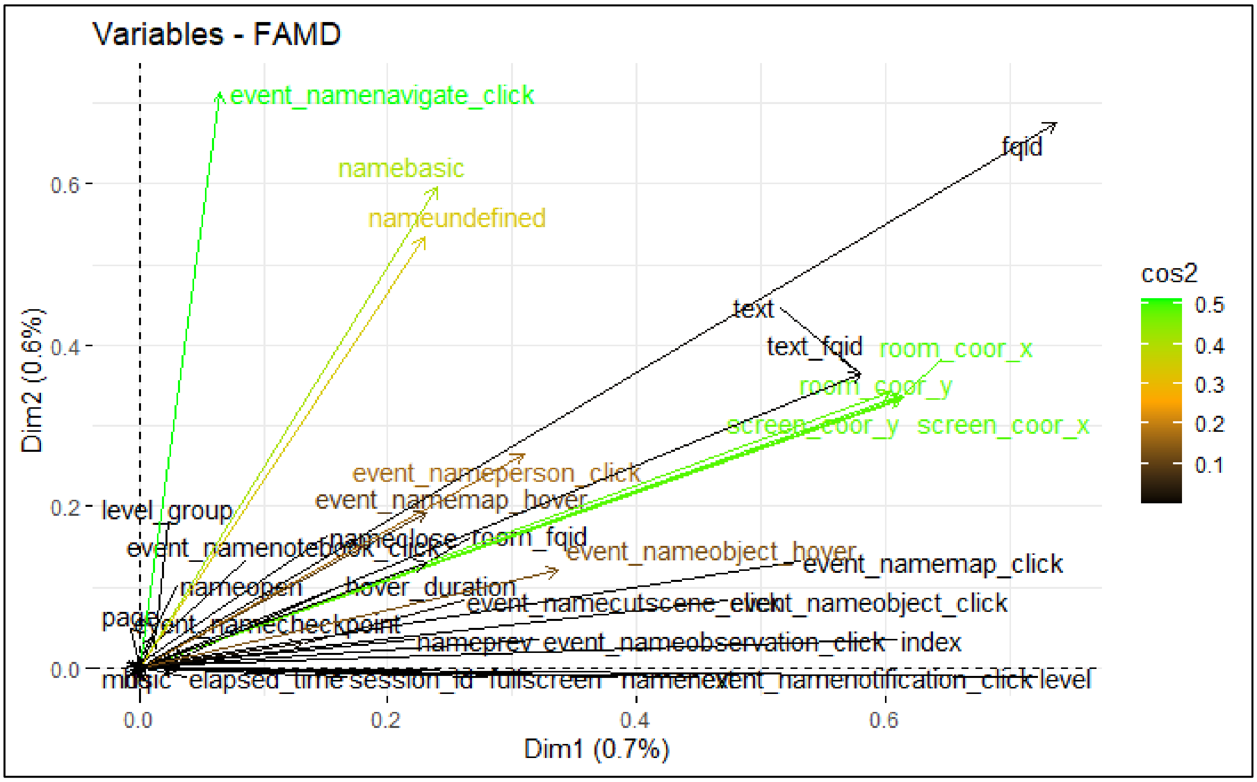


Figure 8. Results biplot +cos2 scores (R)

As we needed to predict the performance of each session_id in 18 different questions, we assumed that the overall prediction accuracy could be maximized by building models separately for each level group. This allowed for more flexibility in inherent changes within these level groups compared to a general model.

Furthermore, as the training data set (train.csv) did not include the response variable ('correct' column), we had to pull this information from another table called train_labels.csv to know if the student was passing the question or not.

Additionally, we defined the level groups differentiating the data points into levels of 0-4, 5-12, and 13-22. They can be directly assigned to questions 1-5, 6-13, and 14-18 (Note: there are only 18 questions, despite having 22 levels). To save memory, we split the level groups into three different files.

Then, after splitting by level_group, we trained the models for each group on the respective training data set. As our response is binary, we needed to use binary classifiers. So we worked with Logistic Regression, Neural Networks, Random Forest, Naïve Bayes, and Tree Learning (XGBoost). Consequently, we ran each model for the three groups and did the same for each question.

Finally, we combined and altered our independent variables with feature engineering. We performed an iterative approach to improving the models by tuning hyperparameters and feature engineering, analyzing each factor's correlation to ensure our independent variables are not too highly correlated. Due to the high number of coefficients and models, we essentially focus on model results from the R², the error measures, and the F1 score. After computing model performance, we compared each measure for each model.

Challenges

We faced challenges in each phase of our analysis. A key challenge was the large amount of data (4.74 GB) which required managing memory resources, efficient coding methods to store and clear memory and creating checkpoints in the data to reduce run times when making changes. It impacted decisions on the number of variables as well as model parameter selection. For example, creating eighteen datasets and eighteen models was extremely memory intensive, resulting in us initially splitting the code file into three files, each for a level grouping. In our first attempts, we could not use the random forest model due to memory constraints. Still, after continued feature engineering and memory resourcefulness and allocation, we successfully ran that specific optimized model type for all eighteen datasets.

Another example of computationally intensive calculations was the FAMD analysis. Performing FAMD on the entire dataset was not practical due to the ~27 million rows which required too large resources in terms of memory. Instead, FAMD was used on a subset of the provided train dataset to look at parameters for feature selection and parameter analysis.

After hot encoding and exploratory data analysis, further challenges appeared in selecting the variables. Many of the variables had low to no predicting power. For example, we attempted a linear regression model to predict a session ID's individual score; this score was taken by how many questions were correctly answered and divided by how many questions in total per session ID. After running the linear regression model, the model results were not significant. We believe this was due to a compounding error; our model tried to predict all 18 results. Indeed, it is worth noting that a classification model seemed more appropriate for the data due to the response being a 0 or 1 for each question.

Finally, we faced a time constraint issue when running the codes. Random Forest takes 50 minutes to run for only one question, indicating a total of 4 hours just to run group 0, so we are not including Random Forest in the html files. However, based on our estimations, running all the models per question would take around 17 hours. Thus, we tested each model individually and generated one file for the best model overall for the groups and questions.

In investigating the test data set, it became clear that no labels were provided for the dataset, as it is part of a Kaggle competition. A requirement of the Kaggle competition is to submit test data runs in one single file, where Kaggle will determine a result based on the accuracy of the predictions. So, instead, we decided to split our train data set into train and test for this project, with respectively 80% and 20%. As the train data set is quite large (around 27,000,000 rows), there was no concern about sample sizes after the split.

RESULTS & EVALUATION

Model Performance

We generated 12 working models in total, four model types for each level group. Random forest was not tested on these datasets as it did not return useable results due to time and memory constraints. R^2, Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and F1 score were used as measures to compare model performance. An overview of all outcomes can be seen in Table 1 which represents the averaged score the different group models achieved. We see in the R^2 values that the models are only able to explain a relatively low portion of the variability in the model. Looking at the RMSEs, it would indicate that we tend to have a relatively low accuracy from our models which can be explained by the large data sets. However, our MAE would suggest that the average error of our models' prediction is not a good value considering that the response is between 0 and 1. The F1 score is lower than expected indicating low precision and recall. Overall, we believe that this approach is generating a pretty low performance.

General Scores	R²	RMSE	MAE	F1 score
Naïve Bayes	0.0001161152	0.6534374	0.4302406	0.1558026
Logistic Regression	0.0000087732	0.6873927	0.4725087	0.004725739
XGBoost	0.0002661065	0.5558103	0.3089251	0.006274023
Neural Net	NA	0.5557179	0.3088223	NA

Table 1 - Overview of Key Metrics per Model for level group model approach – Average Scores

In Table 2 to 4 we see that for every model except Naïve Bayes, the results are the same and that XGBoost has a significantly better R², MAE, and RMSE score in all cases. Random forest was not tested on these datasets due to time and memory constraints. It's important to note that all of the models' F1 score values are relatively low, which suggests that they might not be very good at accurately identifying the positive and negative classes. It is difficult to fairly compare the neural net model to the other models since it lacks the R² and F1 score data.

Group 0	R²	RMSE	MAE	F1 score
Naïve Bayes	0.0002635424	0.5728264	0.3281301	0.109859
Logistic Regression	0.0000087732	0.6873927	0.4725087	0.004725739
XGBoost	0.0002661065	0.5558103	0.3089251	0.006274023
Neural Net	NA	0.5557179	0.3088223	NA

Table 2 - Overview of Key Metrics per Model for level group model approach – Group 0

Group 5	R²	RMSE	MAE	F1 score
Naïve Bayes	0.0000232785	0.6896663	0.4756396	0.1399714
Logistic Regression	0.0000087732	0.6873927	0.4725087	0.004725739
XGBoost	0.0002661065	0.5558103	0.3089251	0.006274023
Neural Net	NA	0.5557179	0.3088223	NA

Table 3 - Overview of Key Metrics per Model for level group model approach – Group 5

Group 13	R²	RMSE	MAE	F1 score
Naïve Bayes	0.0000615247	0.6978195	0.4869520	0.2175906
Logistic Regression	0.0000087732	0.6873927	0.4725087	0.004725739
XGBoost	0.0002661065	0.5558103	0.3089251	0.006274023
Neural Net	NA	0.5557179	0.3088223	NA

Table 4 - Overview of Key Metrics per Model for level group model approach – Group 13

To improve the performance, we took another approach by creating models specifically per question instead of just the level group. We used our initial dataset split by a question and ran logistic regression, Naïve Bayes, XGBoost, and neural net to get the results summarized as averages in Table 5. The complete overview of how each model is performed in detail can be found in Appendix C. This method ran significantly longer and got slightly better MAE and RMSE scores.

	R²	RMSE	MAE	F1 score
Naïve Bayes	0.00428006	0.62102490	0.40926150	0.27154090
Logistic Regression	0.00683485	0.50441310	0.27433250	0.21240810
XGBoost	0.01315661	0.49957180	0.26843720	0.22733460
Neural Net	0.01182713	0.50393030	0.27360600	0.33357590

Table 5 - Overview of Key Metrics per Model for question specific model – Average Scores

Here we see that the performance of models per question that R² was significantly increased by almost one hundred times for each question. However, the maximum R² we could obtain was 1.32% with XGBoost. We also see that RMSE improved with a decrease of 5% and MAE also improved with a decrease of 4% across all models, except for Logistic Regression (decrease of 20% in the MAE) when modelling per question. F1 score significantly increased as well across all models, with Neural Net having the highest F1 score. Consequently, it was a better approach to model at a per question level. Despite the increase in performance indicators, we still do not believe this model is usable due to factors such as computational and time restraints, low overall accuracy, and the method used to hardcode the data.

Evaluation

Based on our initial outcomes, we saw the potential to improve the models through hyperparameter optimization: As XGBoost returned the highest R² value in the initial analysis, we looked at boosting XGBoost performance. Therefore, we changed the booster from gbtree (gradient-boosted tree) to dart (this is also a gradient-boosted tree-based algorithm, however, it drops trees during each round of boosting). Dart tends to lend itself better to preventing overfitting. However, using a dart did not significantly impact overall results in this case. Max_depth was increased to account for a large dataset; this adjustment improved the F1 score for the group-specific model's averaged results (0.004 to 0.006). Finally, gamma was adjusted from 0 to 5 with negligible effects; this is also a regularization parameter and is used to counter-act overfitting. The results above already integrate the optimization of the hyperparameters.

Even though the primary measure for our models' performance in the competition is the F1 score, we took the other performance indicators into more significant consideration. Even though Neural Net had the highest F1 score, we moved forward with the XGBoost classifier due to the MAE and RMSE. The low RMSE implies a good fit and the low MAE indicates higher accuracy. When running our models per level group, we still see that XGBoost was the best model because of performance indicators outside of the F1 score, though Naïve Bayes had the best F1 score. We decided to value accuracy and performance indicators in this case when choosing between model per question and model per level group due to the promising performance indicators outside of the F1 score.

In general, we see that the performance of all models was low. It would strongly indicate that the used variables have too low predicting power to make reasonable predictions. Also, the improved models could not solve this challenge and further optimization of the given variables will most likely not significantly change this. Instead, we propose considering additional variables to improve the prediction.

Our next and final step is to generate our code file with this model. The code file will contain the eighteen XGBoost models (one model per question). Although, as mentioned above, we decided to move forward with models per question due to the significant increase in performance indicators, even though the model takes more time to run, we believe that the model's accuracy should be the final priority. Therefore, our file will initially hot encode the categorical variables from the input data, before running the data into our models. The model will then be able to give predictions per session ID and questions.

CONCLUSION

Business Implications

Our analysis uses a large variety of data to predict whether a question will be answered correctly or not. In the context of business, similar models could potentially be used to predict if a person is likely to make a correct decision under given circumstances. To analyze this, we do not necessarily need to have access to personal information.

Predictive analysis on large and complex datasets with multiple user inputs has beneficial uses across many industries; a similar approach could be used to predict the likelihood of a customer making a purchase based on past purchases and external factors. In this concrete example, we see the benefit of having large amounts of data accessible for analysis. However, in the case of educational games and improving student learning, the work done on this project is a good example of applying predictive modeling methodologies. In this case, we were able to show that external factors alone will, in many cases, not be enough to predict the outcome of tasks.

Outlook

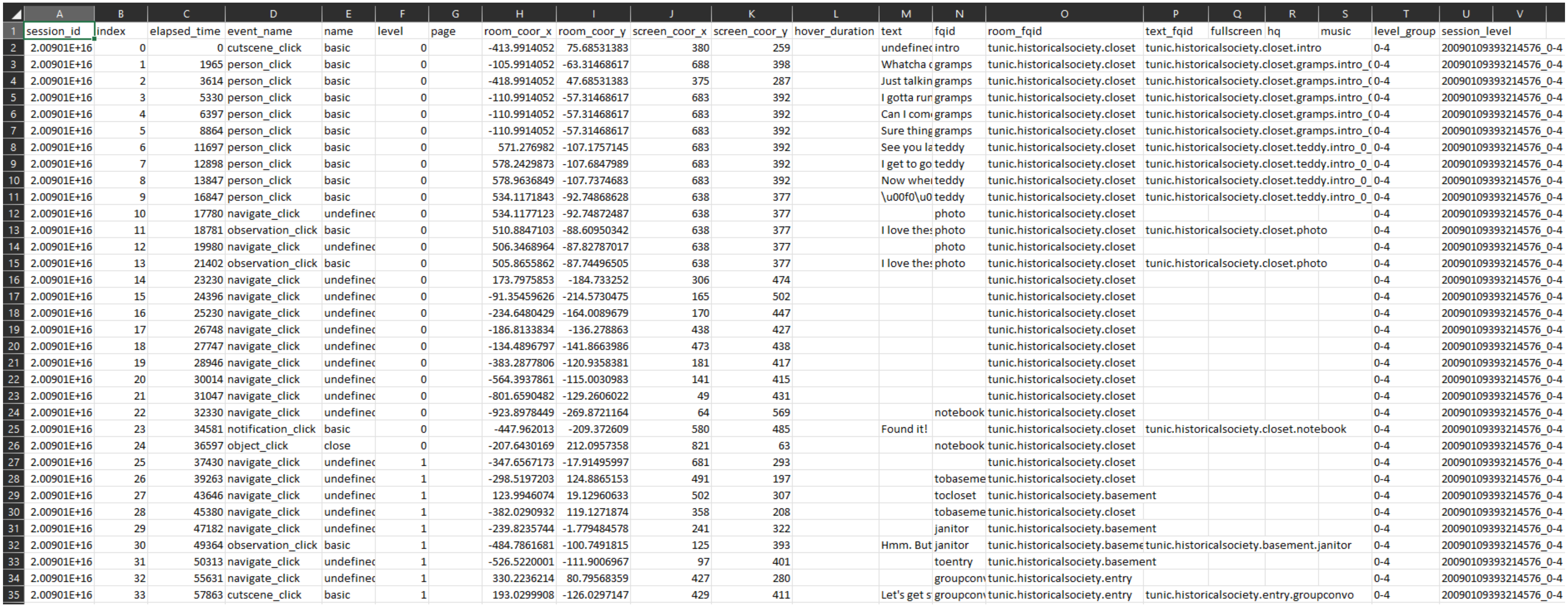
Overall model performance was poor and should not be used for predicting results in the current state. However, this was impacted by computational and time restraints. There would be value in investigating performance with other parameters and feature selection methodologies.

While the initial performance was low, there is good reason to believe there is potential for value with more time and resources. Furthermore, we believe that hyperparameter optimization and additional computational power to handle the extensive data and features would help improve results. In general, we were able to identify variables that have low predicting power and could focus on collecting data related to the more impactful variables. It could also be valuable to combine the data with personal user profiles; a person's past performance is likely a strong indicator of future performance as well. We think there is value in pursuing past performance if time and resources allowed; this would also allow for understanding a student's performance over time. There would be value in considering scaling the variables, utilizing past performance data per individual, as well as including other FAMD dimensions in the analysis, as these were also impacted by the mentioned computational constraints. Further focus on improving the efficiency of the code, reducing the computational limits, and utilizing past performance variables would also allow for better results; model choices, model parameters, and feature selection were all impacted by the intensive requirements to run the dataset.

To conclude this project, evaluating different metrics will result in other models performing better. While there was overall poor model performance, we would recommend focusing on improving XGBoost first due to getting better R^2 and RMSE scores. Afterward, we will pursue the Kaggle competition and attempt to rectify our code with the above recommendations before the deadline on June 7th as the first team posting a full code in R.

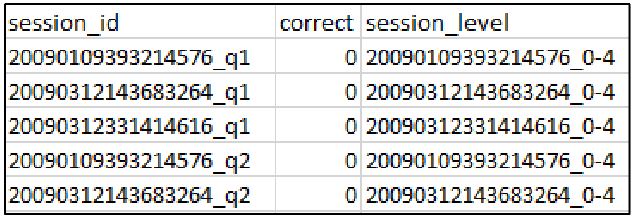
APPENDIX A – Dataset screenshot

Test table



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	session_id	index	elapsed_time	event_name	name	level	page	room_coor_x	room_coor_y	screen_coor_x	screen_coor_y	hover_duration	text	fqid	room_fqid	text_fqid	fullscreen	hq	music	level_group	session_level	
2	2.00901E+16	0	0	cutscene_click	basic	0		-413.9914052	75.68531383	380	259		undefinec	intro	tunic.historicalsociety.closet	tunic.historicalsociety.closet.intro				0-4	20090109393214576_0-4	
3	2.00901E+16	1	1965	person_click	basic	0		-105.9914052	-63.31468617	688	398		Whatcha c	gramps	tunic.historicalsociety.closet	tunic.historicalsociety.closet.gramps.intro_(				0-4	20090109393214576_0-4	
4	2.00901E+16	2	3614	person_click	basic	0		-418.9914052	47.68531383	375	287		Just talkin	gramps	tunic.historicalsociety.closet	tunic.historicalsociety.closet.gramps.intro_(				0-4	20090109393214576_0-4	
5	2.00901E+16	3	5330	person_click	basic	0		-110.9914052	-57.31468617	683	392		I gotta rur	gramps	tunic.historicalsociety.closet	tunic.historicalsociety.closet.gramps.intro_(				0-4	20090109393214576_0-4	
6	2.00901E+16	4	6397	person_click	basic	0		-110.9914052	-57.31468617	683	392		Can I com	gramps	tunic.historicalsociety.closet	tunic.historicalsociety.closet.gramps.intro_(				0-4	20090109393214576_0-4	
7	2.00901E+16	5	8864	person_click	basic	0		-110.9914052	-57.31468617	683	392		Sure thing	gramps	tunic.historicalsociety.closet	tunic.historicalsociety.closet.gramps.intro_(				0-4	20090109393214576_0-4	
8	2.00901E+16	6	11697	person_click	basic	0		571.276982	-107.1757145	683	392		See you la	teddy	tunic.historicalsociety.closet	tunic.historicalsociety.closet.teddy.intro_0_				0-4	20090109393214576_0-4	
9	2.00901E+16	7	12898	person_click	basic	0		578.2429873	-107.6847989	683	392		I get to go	teddy	tunic.historicalsociety.closet	tunic.historicalsociety.closet.teddy.intro_0_				0-4	20090109393214576_0-4	
10	2.00901E+16	8	13847	person_click	basic	0		578.9636849	-107.7374683	683	392		Now wher	teddy	tunic.historicalsociety.closet	tunic.historicalsociety.closet.teddy.intro_0_				0-4	20090109393214576_0-4	
11	2.00901E+16	9	16847	person_click	basic	0		534.1171843	-92.74868628	638	377		\u00f0\u0	teddy	tunic.historicalsociety.closet	tunic.historicalsociety.closet.teddy.intro_0_				0-4	20090109393214576_0-4	
12	2.00901E+16	10	17780	navigate_click	undefinec	0		534.1177123	-92.74872487	638	377			photo	tunic.historicalsociety.closet					0-4	20090109393214576_0-4	
13	2.00901E+16	11	18781	observation_click	basic	0		510.8847103	-88.60950342	638	377		I love thes	photo	tunic.historicalsociety.closet	tunic.historicalsociety.closet.photo				0-4	20090109393214576_0-4	
14	2.00901E+16	12	19980	navigate_click	undefinec	0		506.3468964	-87.82787017	638	377			photo	tunic.historicalsociety.closet					0-4	20090109393214576_0-4	
15	2.00901E+16	13	21402	observation_click	basic	0		505.8655862	-87.74496505	638	377		I love thes	photo	tunic.historicalsociety.closet	tunic.historicalsociety.closet.photo				0-4	20090109393214576_0-4	
16	2.00901E+16	14	23230	navigate_click	undefinec	0		173.7975853	-184.733252	306	474				tunic.historicalsociety.closet					0-4	20090109393214576_0-4	
17	2.00901E+16	15	24396	navigate_click	undefinec	0		-91.35459626	-214.5730475	165	502				tunic.historicalsociety.closet					0-4	20090109393214576_0-4	
18	2.00901E+16	16	25230	navigate_click	undefinec	0		-234.6480429	-164.0089679	170	447				tunic.historicalsociety.closet					0-4	20090109393214576_0-4	
19	2.00901E+16	17	26748	navigate_click	undefinec	0		-186.8133834	-136.278863	438	427				tunic.historicalsociety.closet					0-4	20090109393214576_0-4	
20	2.00901E+16	18	27747	navigate_click	undefinec	0		-134.4896797	-141.8663986	473	438				tunic.historicalsociety.closet					0-4	20090109393214576_0-4	
21	2.00901E+16	19	28946	navigate_click	undefinec	0		-383.2877806	-120.9358381	181	417				tunic.historicalsociety.closet					0-4	20090109393214576_0-4	
22	2.00901E+16	20	30014	navigate_click	undefinec	0		-564.3937861	-115.0030983	141	415				tunic.historicalsociety.closet					0-4	20090109393214576_0-4	
23	2.00901E+16	21	31047	navigate_click	undefinec	0		-801.6590482	-129.2606022	49	431				tunic.historicalsociety.closet					0-4	20090109393214576_0-4	
24	2.00901E+16	22	32330	navigate_click	undefinec	0		-923.8978449	-269.8721164	64	569			notebook	tunic.historicalsociety.closet					0-4	20090109393214576_0-4	
25	2.00901E+16	23	34581	notification_click	basic	0		-447.962013	-209.372609	580	485		Found it!		tunic.historicalsociety.closet	tunic.historicalsociety.closet.notebook				0-4	20090109393214576_0-4	
26	2.00901E+16	24	36597	object_click	close	0		-207.6430169	212.0957358	821	63			notebook	tunic.historicalsociety.closet					0-4	20090109393214576_0-4	
27	2.00901E+16	25	37430	navigate_click	undefinec	1		-347.6567173	-17.91495997	681	293				tunic.historicalsociety.closet					0-4	20090109393214576_0-4	
28	2.00901E+16	26	39263	navigate_click	undefinec	1		-298.5197203	124.8865153	491	197			tobaseme	tunic.historicalsociety.closet					0-4	20090109393214576_0-4	
29	2.00901E+16	27	43646	navigate_click	undefinec	1		123.9946074	19.12960633	502	307			tocloset	tunic.historicalsociety.basement					0-4	20090109393214576_0-4	
30	2.00901E+16	28	45380	navigate_click	undefinec	1		-382.0290932	119.1271874	358	208			tobaseme	tunic.historicalsociety.closet					0-4	20090109393214576_0-4	
31	2.00901E+16	29	47182	navigate_click	undefinec	1		-239.8235744	-1.779484578	241	322			janitor	tunic.historicalsociety.basement					0-4	20090109393214576_0-4	
32	2.00901E+16	30	49364	observation_click	basic	1		-484.7861681	-100.7491815	125	393		Hmm. But	janitor	tunic.historicalsociety.baseme	tunic.historicalsociety.basement.janitor				0-4	20090109393214576_0-4	
33	2.00901E+16	31	50313	navigate_click	undefinec	1		-526.5220001	-111.9006967	97	401			toentry	tunic.historicalsociety.basement					0-4	20090109393214576_0-4	
34	2.00901E+16	32	55631	navigate_click	undefinec	1		330.2236214	80.79568359	427	280			groupcon	tunic.historicalsociety.entry					0-4	20090109393214576_0-4	
35	2.00901E+16	33	57863	cutscene_click	basic	1		193.0299908	-126.0297147	429	411		Let's get s	groupcon	tunic.historicalsociety.entry	tunic.historicalsociety.entry.groupconvo				0-4	20090109393214576_0-4	

Sample submission table



session_id	correct	session_level
20090109393214576_q1	0	20090109393214576_0-4
20090312143683264_q1	0	20090312143683264_0-4
20090312331414616_q1	0	20090312331414616_0-4
20090109393214576_q2	0	20090109393214576_0-4
20090312143683264_q2	0	20090312143683264_0-4

APPENDIX B – Columns' description

Columns	Description
session_id	the session ID in which the event occurred
index	the event index for the session
elapsed_time	how much time has elapsed (in milliseconds) between the session start and the recording event
event_name	the event type's name
name	the event name (ie: whether a notebook click opens or closes the notebook)
level	what game level the event happened in (0 to 22)
page	the event's page number (only for notebook-related events)
room_coor_x	the click coordinates related to the in-game room (only for click events)
room_coor_y	the click coordinates related to the in-game room (only for click events)
screen_coor_x	the click coordinates related to the player's screen (only for click events)
screen_coor_y	the click coordinates related to the player's screen (only for click events)
hover_duration	how long the hover lasted (in milliseconds) (only for hovering events)
text	the text displayed to the player during this event
fqid	the event's fully qualified ID
room_fqid	the fully qualified ID of the room where the incident occurred
text_fqid	the fully qualified ID of the text where the incident occurred
fullscreen	whether or not the player is running in fullscreen mode
hq	whether or not the player is running in high-quality mode
music	whether or not the player is playing some music
level_group	which group of levels - and group of questions - this row belongs to (0-4, 5-12, 13-22)

APPENDIX C – Models' results per question

- **Logistic Regression**

Question	R2	RMSE	MAE	f1_score
1	2.19E-03	0.545408	0.29747	2.99E-02
2	1.02E-06	0.159813	0.02554	NaN
3	6.80E-06	0.277418	0.07696	6.53E-05
4	5.06E-03	0.472049	0.222831	3.18E-02
5	3.77E-02	0.637019	0.405793	5.20E-01
6	4.67E-04	0.513183	0.263356	1.33E-04
7	6.93E-03	0.542825	0.294659	7.43E-02
8	5.74E-04	0.636271	0.404841	NaN
9	1.09E-02	0.543503	0.295395	9.96E-02
10	1.59E-03	0.68042	0.462971	6.69E-01
11	8.17E-04	0.622419	0.387405	NaN
12	7.48E-03	0.396433	0.157159	4.32E-02
13	3.08E-05	0.506849	0.256896	8.53E-01
14	2.01E-02	0.563319	0.317328	1.80E-01
15	2.88E-02	0.634011	0.40197	6.68E-01
16	2.87E-04	0.52772	0.278489	8.71E-03
17	7.33E-05	0.571762	0.326912	8.13E-03
18	3.43E-06	0.249016	0.062009	8.73E-04

- **Neural Net**

Question	R2	RMSE	MAE	f1_score
1	NA	0.54583	0.29793	NA
2	NA	0.159691	0.025501	NA
3	NA	0.277175	0.076826	NA
4	4.62E-03	0.472091	0.22287	2.75E-02
5	2.96E-02	0.643515	0.414112	5.90E-01
6	NA	0.511809	0.261948	NA
7	3.26E-09	0.544978	0.297001	3.81E-05
8	NA	0.635606	0.403995	NA
9	4.02E-03	0.545841	0.297943	1.76E-02
10	2.42E-02	0.646518	0.417986	6.26E-01
11	9.94E-03	0.616041	0.379506	2.15E-01
12	NA	0.397221	0.157785	NA
13	NA	0.505922	0.255957	8.53E-01
14	2.55E-03	0.572532	0.327793	4.42E-02
15	1.97E-02	0.649698	0.422107	6.29E-01
16	NA	0.527266	0.27801	NA
17	NA	0.571138	0.326199	NA
18	NA	0.247875	0.061442	NA

- **XGBoost**

Question	R2	RMSE	MAE	f1_score
1	1.06E-02	0.54145	0.293168	0.087454
2	1.18E-03	0.159596	0.025471	0.002852
3	3.34E-03	0.276681	0.076552	0.009168
4	5.73E-03	0.471529	0.22234	0.026677
5	4.01E-02	0.632439	0.399979	0.57062
6	1.62E-02	0.507543	0.2576	0.125619
7	8.75E-03	0.541481	0.293202	0.076474
8	6.55E-03	0.630024	0.39693	0.154966
9	1.56E-02	0.541397	0.29311	0.128661
10	2.85E-02	0.64178	0.411882	0.627629
11	1.55E-02	0.610907	0.373207	0.246699
12	8.47E-03	0.395184	0.156171	0.028924
13	6.19E-05	0.5059	0.255935	0.853245
14	2.67E-02	0.561003	0.314724	0.237146
15	3.37E-02	0.630319	0.397302	0.663614
16	1.09E-03	0.526821	0.27754	0.007499
17	1.56E-03	0.570364	0.325315	0.01744
18	NA	0.247875	0.061442	NA

- **Naives Bayes**

Question	R2	RMSE	MAE	f1_score
1	1.27E-03	0.551214	0.303837	0.067574
2	3.75E-05	0.958987	0.919656	0.050123
3	1.11E-03	0.311764	0.097197	0.068378
4	2.63E-03	0.48434	0.234585	0.086902
5	1.82E-02	0.67059	0.449691	0.289724
6	6.41E-04	0.652286	0.425477	0.322436
7	1.00E-02	0.54537	0.297428	0.143948
8	2.89E-04	0.753804	0.56822	0.536546
9	1.18E-02	0.549538	0.301992	0.178256
10	5.32E-05	0.688548	0.474099	0.676176
11	1.03E-03	0.775613	0.601576	0.53436
12	1.17E-03	0.773094	0.597674	0.269142
13	6.53E-03	0.746578	0.557379	0.473854
14	1.17E-02	0.581318	0.33793	0.255606
15	7.43E-03	0.666906	0.444764	0.621467
16	2.27E-03	0.534259	0.285433	0.0858
17	5.42E-06	0.596286	0.355557	0.137878
18	8.22E-04	0.337954	0.114213	0.089567

REFERENCES

Franklin, A., Gagnon, D., HCL-Jevster, Maggie, Benner, M., Rambis, N., . . . Boser, U. (2023). *Predict Student Performance from Game Play*. Retrieved from Kaggle: https://kaggle.com/competitions/predict-student-performance-from-game-play

Manzano-León, A., Camacho-Lazarraga, P., Guerrero, M. A., Guerrero-Puerta, L., Aguilar-Parra, J. M., Trigueros, R., & Alias, A. (2021). Between Level Up and Game Over: A Systematic Literature Review of Gamification in Education. *Sustainability, 13*(1), 1-14. doi:doi.org/10.3390/su13042247

Nietfeld, J. L. (2020). Predicting transfer from a game-based learning environment. *Computers & Education, 1*(1), 103780. doi:doi.org/10.1016/j.compedu.2019.103780

Visbal-Cadavid, D., Mendoza-Mendoza, A., & De La Hoz-Dominguez, E. (2020). Use of Factorial Analysis of Mixed Data (FAMD) and Hierarchical Cluster Analysis on Principal Component (HCPC) for Multivariate Analysis of Academic Performance of Industrial Engineering Programs. *Journal of Southwest Jiaotong University, 55*(5), 1-16. doi:doi.org/10.35741/issn.0258-2724.55.5.34

Wang, L.-H., Chen, B., Hwang, G.-J., Guan, J.-Q., & Wang, Y.-Q. (2022). Effects of digital game-based STEM education on students' learning achievement a meta-analysis. *International Journal of STEM Education, 9*(26), 1-13. doi:doi.org/10.1186/s40594-022-00344-0