Utilizing Machine Learning Methods to Predict Student Re-engagement via Minecraft Data

Mylene Villegas
Department of Information
Systems and Computer Science
Ateneo de Manila University
Quezon City, Philippines
mylenebvillegas@gmail.com

ABSTRACT

This study focused on developing a model to predict student re-engagement (a student's voluntary re-participation in an educational intervention) using the data extracted from a Game-Based Learning called What-if Hypothetical Implementations in Minecraft (WHIMC), a game which offers engaging STEM learning experiences through Minecraft Java Edition. Quantifying re-engagement in environments like Minecraft can be considerably difficult because of its open-ended structure. In order to address this, the researcher harnessed machine learning methods to deliver a more advanced and thorough result, going beyond the traditional approaches of previous studies. The current research utilized data gathered from a group of Grade 8 students at a middle school in the Philippines who engaged with WHIMC. This collection of data includes student demographics, survey responses, and in-game features such as player positions, observations, and science tools usage. Leveraging the collected dataset, various machine learning classifiers, including Logistic Regression, Random Forest, Naïve Bayes, Support Vector Machine (SVM), and Multilayer Perceptron (MLP) algorithms, were evaluated using metrics such as ROC-AUC, Sensitivity, and Specificity. Notably, the MLP classifier, utilizing both survey and in-game data, exhibited the most optimal performance, yielding an ROC-AUC of .81, Sensitivity of 71%, and Specificity of 79%. Highlighting the significance of the Specificity metric, this study underscores its role in directing interventions towards students with a higher likelihood of disengagement.

KEYWORDS

Minecraft, whimc, machine learning, student re-engagement

1 INTRODUCTION

Re-engagement refers to a learner's spontaneous and self-motivated re-participation in an educational intervention. It is important as it serves as evidence of intrinsic motivation and intrinsic motivation leads students to demonstrate enhanced learning effectiveness [8, 9, 38]. This is one of the phases of the engagement process model [21]. This model is focused on deconstructing the concept of engagement concerning individuals' experiences with technology. It is composed of four distinct phases: the point of initial engagement, a phase of sustained engagement, a subsequent disengagement phase, and, in some cases, a potential re-engagement phase.

Multiple studies indicate that when students engage in specific educational activities, they generally experience higher levels of satisfaction, better academic performance, persistence, and social engagement [2, 15, 16]. The level of student engagement can be impacted by various contextual factors, such as the specific learning environments they find themselves in or the instructional approaches employed by their teachers [6]. These external elements play a significant role in shaping how actively students participate in their educational experiences.

In parallel with this heightened interest on student engagement, the integration of digital technology has emerged as a central component of higher education, exerting its influence across every facet of the student experience [4, 12, 27]. A prominent example of the utilization of digital technology for learning is the integration of game-based learning (GBL). Implementing GBL offered students engaging learning experiences through interactive means. This approach granted students autonomy over tasks, amplifying their curiosity and involvement while preventing unfavorable emotional encounters [31].

This study focused on predicting student re-engagement using the data extracted from a GBL called What-if Hypothetical Implementations in Minecraft (WHIMC). WHIMC utilizes Minecraft Java Edition as an educational platform for students to actively delve into the scientific implications of alternate renditions of Earth via "what if" questions, such as "What if the Earth had no moon?" or "What if the Earth has a colder sun?" (As depicted in Figure 1). The project's primary goal is to engage, excite, and generate interest in STEM (Science, Technology, Engineering, and Mathematics) [35]. The data extracted from the game includes in-game data, student demographics, and answers to survey questions, which contain numerical, categorical, ordinal, and free-text data coming from the students who took the module.

Figure 1: Earth with a Colder Sun in WHIMC. The world exhibits two distinct regions: the Western Area (left), characterized by a desert landscape, and the Eastern Area (right), covered in snow.

GBLs have proven to be beneficial for education across various fields, offering advantages such as fostering more genuine learning experiences and boosting student engagement. This is primarily due to their high level of interactivity and immersion [1]. In addition to this, the favorable impacts of GBLs encompass multiple areas such as student performance, skill acquisition, enthusiasm, and involvement. This motivated educators, game developers, funding entities, and researchers to employ games for instructing STEM subjects on various platforms [5]. Similar to these, WHIMC, as a GBL, shares the goal of engaging, nurturing enthusiasm, and sparking curiosity within the realm of science education [35]. As students immerse themselves in GBLs and other educational technologies like WHIMC, their learning potential expands. While re-engagement is not obligatory, it remains valuable in facilitating the learning opportunities and benefits mentioned in this chapter.

However, measuring re-engagement in environments like Minecraft can be considerably difficult because of its open-ended structure. For instance, the work of [37] tried to do an analysis on the impacts of WHIMC on student interest by making use of both quantitative and qualitative data such as coded interviews and notes, surveys on STEM interest, assessments of knowledge, and self-reported skill levels in Minecraft. On the other hand, the study by [9] attempted to quantify re-engagement by making use of additional in-game data on top of surveys, such as player positions, player observations and Science tools usage logs. Another study by [33] asked parents for help to check if students were playing Minecraft at home and to observe the students' level of concentration during their interaction with the game.

In the current research endeavor, the researcher attempted to quantify re-engagement, transcending the conventional approaches previously outlined and instead, utilizing machine learning methods, which have the potential to deliver a more sophisticated and comprehensive outcome. Through this study, the researcher investigated the following: **RQ 1:** How can we develop a machine learning pipeline to effectively predict a student's likelihood of re-engaging with WHIMC? **RQ 2:** How effectively does the best-performing machine learning model predict re-engagement?

The rest of this paper is organized as follows. Included in the next section is the discussion of literature that provides background on the significance and potential benefits of utilizing machine learning methods to create a model that will be able to predict re-engagement using the survey and in-game data derived from WHIMC. This is then followed by a section that focuses on methods that the author used for data pre-processing, feature engineering, feature selection, modeling, validation, and post-modeling analysis. After, the results are shown by comparing the results of the various models. Finally, the author presents the conclusions and identify directions for future research.

2 REVIEW OF RELATED LITERATURE

To establish the groundwork for this research endeavor, the researcher conducted the literature review, thoroughly examining the existing body of work that employed machine learning techniques to quantify engagement across various platforms.

2.1. Existing studies about predicting engagement

Domina et al. [11] employed Machine Learning (ML) to predict student engagement during the COVID-19 pandemic, analyzing data from a survey of 10,000 parents in a southeastern U.S. school district. The resulting regression models highlighted the significance of technology accessibility: students with high-speed internet and devices showed greater engagement, even after socioeconomic factors were considered. Access to diverse learning opportunities also correlated with higher engagement levels. This discovery emphasizes that educators, despite their limited control over students' family backgrounds and household resources, retain the ability to make critical decisions regarding the instructional materials and resources they provide [11].

A different research conducted by Hussain et al. [13] employed machine learning algorithms to identify students with low engagement levels in a social science course from a virtual learning environment (VLE) at the Open University (OU) in the United Kingdom. To identify low-engagement students, the authors applied several types of ML algorithms to the dataset. They initially trained models using these algorithms and subsequently compared their accuracy, kappa values, and recall rates. The results indicated that the J48, decision tree, JRIP, and gradient-boosted classifiers demonstrated good performance in terms of accuracy, kappa value, and recall compared to other models. Based on these findings, the authors developed a dashboard to assist instructors at OU. These models can be seamlessly integrated into VLE systems, allowing instructors to assess student engagement across different activities

and materials and offer timely interventions to support students before their final exams. [13]

Another literature in line with the previously enumerated studies is the work by Ayouni et al. [3] that proposes an intelligent predictive system capable of forecasting students' engagement levels and providing them with feedback to enhance their motivation and commitment. A variety of data points such as number of logins, user participation, user activity, etc. were extracted from a Learning Management System (LMS). Based on these input features, the authors aim to create a predictor that could categorize students into three groups based on their engagement levels: "Not Engaged," "Passively Engaged," and "Actively Engaged." Three different machine learning algorithms, specifically the Decision Tree, Support Vector Machine, and Artificial Neural Network, were applied to analyze students' activities recorded in LMS reports. [3]

The findings of the study indicate that ML algorithms are effective in predicting students' engagement levels. Among these algorithms, the Artificial Neural Network displayed the highest accuracy rate at 85%, surpassing the Support Vector Machine (80%) and Decision Tree (75%) classification techniques. Based on these results, the intelligent predictive system delivers feedback to students and alerts instructors when a student's engagement level decreases. Instructors can then identify students' challenges within the course and motivate them through means such as email reminders, course messages, or scheduling online meetings. [3]

2.2 Attempts to quantify re-engagement

As of writing, there is only one study that aimed to quantify student re-engagement in the context of an open-world GBL. The study by Casano et al. [9] was initialized by gathering data through a survey to identify the most preferred features of WHIMC, followed by a systematic coding of these features to determine which aspects could be quantified and analyzed. Three potential re-engagement triggers were identified: social play, free exploration, and interactive learning components within WHIMC. The authors then used in-game data such as player position records, observations, and Science tool usage to develop simple heuristics to identify whether these triggers are exhibited outside designated testing hours. For instance, social play was defined as instances where player positions overlapped, indicating concurrent playtime outside of testing hours. The authors were able to utilize the module's design, as students had the option to access WHIMC beyond testing hours, enabling the analysis to consider these interactions as voluntary and unprompted. [9]

Furthermore, the study demonstrated ways to establish the extent to which specific elements may have contributed to re-engagement. For example, the degree of social play as a potential re-engagement trigger was described using the average number of concurrent users outside of testing hours. Free exploration as a re-engagement element was described by examining the ratio between the worlds visited and the time spent exploring outside testing hours. Finally, the study illustrated how a creative parsing of available datasets enabled an understanding of the extent to which the interactive elements in WHIMC played a role in students' re-engagement with the game. [9]

3 METHODOLOGY

This section will highlight the methods employed by the researcher for data collection, data description, data cleaning, feature engineering, feature selection, modeling, and post-modeling analysis. It essentially outlines the approach and techniques used to address RQ1 and RQ2. Note that all these methodologies will be implemented using the Python programming language. Providing context for the collected data in the study, this section begins with details on navigating the WHIMC Server.

3.1 Navigating the WHIMC Server

WHIMC is a set of Minecraft worlds that aims to make learning STEM fun and engaging. These worlds are created and deployed in Minecraft Java Edition, and students can explore them to learn about different topics on STEM subjects. As the players progress in the game, they will have the opportunity to teleport to multiple worlds such as various versions of Earth which include normal baseline Earth, Earth with no moon, Earth with a different tilt, and Earth at a much cooler temperature. For instance, "Earth with No Moon" serves as an introductory experience for players as they begin exploring "what-if" scenarios. In this simplified alternate Earth, players gradually learn about various scientific concepts and tools, with a special emphasis on the observation tool, which is a crucial part of navigating the server. Completing this world equips players with a solid foundation for approaching other "what-if" scenarios more confidently. The objectives for this experience include gaining insights into how the absence of a moon impacts Earth's tides and winds, understanding the relationship between green energy sources and specific environmental conditions, and becoming proficient in using fundamental science variables and the /observe command—skills vital for success on the server. Similar to the “Earth with No Moon” world, every other world within the WHIMC server is thoughtfully designed to offer a diverse array of activities and valuable learning opportunities that cater to the players' educational and exploratory needs. [36]

3.2 Data Collection

The data used for this study came from a collective effort of Tablatin, Casano, and Rodrigo for previous studies focusing on WHIMC [9, 30]. The research team partnered with Philippine schools and teachers to develop learning modules targeting parts of the curriculum where WHIMC could be integrated. The data that was utilized for [30] was collected from the entire Grade 8 school population consisting of 8 class sections of a middle school in the Philippines. The students were tasked to work on two learning

modules, with ecosystem as the theme for Module 1 and biodiversity and evolution for Module 2. Prior to utilizing WHIMC, students completed a pre-test called Stem Interest Questionnaire (SIQ) to gauge their interest in various subject areas. Subsequently, following their engagement with WHIMC, the students participated in knowledge assessments, the Game Experience Questionnaire (GEQ), and the post-SIQ as a post-test evaluation. The respondents rate their level of agreement for both SIQ and GEQ using a 5-point Likert scale format (1 – *strongly disagree/not at all*, 2 – *disagree/slightly*, 3 – *neutral/moderately*, 4 – *agree/fairly*, 5 – *strongly agree/extremely*). After the activities mentioned, the researchers were able to come up with a collection of data which includes in-game features such as player positions, observations, and science tools used. In addition to this, student demographics and survey responses which includes their responses to the SIQ, GIQ, and open-ended questions were also compiled. In the current study, the researcher consolidated these data and performed the necessary steps enumerated in the next sections.

3.3 Consolidated Data Overview

In total, there were 211 prospective participants from the Grade 8 school population, representing 8 class sections of a middle school in the Philippines. After refining the data to include only those students who completed all the surveys and actively participated in WHIMC gameplay, 116 students, aged 13-15, remained. Among these 116 participants, a significant majority (93 out of 116 or 80%) reported being either familiar or very familiar with Minecraft. Another 20 students (17%) considered themselves somewhat familiar, while only 3 students (3%) admitted to having no prior familiarity.

A custom WHIMC plugin tracked player positions from a top-view perspective within the virtual worlds, capturing an average of 2,867 position counts per student. Additionally, each student generated an average of 10 observations and made use of science tools an average of 11 times during their gameplay.

3.4 Data Cleaning

It is important to acknowledge that the collected data may inherently contain noise, which can have a negative impact on the model's training. Hence, the process of data cleaning was carried out as deemed necessary [10]. In this sub-section, the researcher provides the methodology on how the data was refined, sanitized, and structured for the subsequent steps of the study.

3.4.1 Correcting Data Types

The initial data cleaning phase involved aligning data fields with appropriate types for compatibility with Python libraries. Utilizing functions from the Pandas Library [22], IDs across various dataframes were standardized to string format, ensuring seamless merging. Survey timestamps were transformed into datetime format, while in-game datasets with epoch timestamps were converted to human-readable datetime values using a lambda function.

3.4.2 Handling Missing Values

Since the study is primarily focused on data derived from survey responses, instances (rows) representing students who were unable to complete all the surveys were filtered out across the datasets. On the other hand, the target variable for the model will be based on the instances of the students who will re-engage with WHIMC; therefore, students who did not actually participate in the game were excluded from the datasets. These steps were done through DataFrame indexing and boolean indexing in Pandas Library [22].

3.4.3 Data Standardization

Survey responses, primarily structured on the Likert Scale, required standardization due to varying interpretations. A mapping dictionary was created to flip values in columns with inverse responses. This ensured a consistent representation of the Likert scale, facilitating straightforward data analysis and interpretation.

3.4.4 Textual Data Clean-up

The survey responses were stripped of white spaces, unicode characters, and punctuation marks and lastly converted into lowercase. The Natural Language Toolkit (NLTK) [20] package was then used to remove English stop words.

The textual data is then normalized through lemmatization. Lemmatization is a method employed to reduce words to their base or root form, known as the lemma [14]. It utilizes a dictionary in its implementation. Put simply, it identifies the root word from various text variations. For instance, if the dataset contains words like “changes,” “changed,” and “changing,” all these words will be transformed to their root word, “change.” The Python implementation of this algorithm was accomplished by utilizing functions from NLTK [20].

3.5 Feature Engineering

The results of a closely related study [9] successfully identified three specific elements that emerged as potential drivers for re-engagement: social play, free exploration, and interactive learning elements within WHIMC. These findings were obtained through the systematic coding of in-game features, including player positions, observations, and Science tools usage. This could indicate that the relevant features for predicting student re-engagement may be found within the in-game features. To confirm this, the current study also focused on engineering features from the in-game dataframes.

3.5.1 Categorical Data Encoding

In this study, three techniques were utilized to convert categorical variables into numerical formats: One-Hot Encoding, Ordinal Encoding, and Binary Encoding. One-Hot Encoding was applied to variables without ordinal relationships, transforming them into binary variables indicating the presence (1) or absence (0) of specific categories. Ordinal Encoding preserved the order or ranking of categories, assigning numerical values based on their hierarchy. For example, "Familiarity" categories were encoded from 1 to 5. Binary Encoding captured the presence or absence of specific categorical values, setting binary indicators to 1 if a category was present and 0 otherwise.

3.5.2 In-game Data Aggregation

Aggregation techniques for feature engineering involve creating new features by grouping information from existing features within a dataset. These techniques are common concepts in descriptive statistics, used to capture higher-level patterns and information that can be valuable for machine learning models [24]. In the context of this study, the in-game data includes multiple rows for each participant. To create features that describe the details of the participant's gameplay, the researcher grouped the in-game data at the student level, utilizing standard aggregation functions such as mean, sum, count, minimum, and maximum.

3.6 Additional Features from Open-Ended Survey Responses

Analyzing linguistic data through computational methods is the essence of Natural Language Processing (NLP). Generally, the objective is to construct a representation of the text that imparts structure to the inherent unstructured nature of natural language [34]. NLP methods were employed on the survey responses to engineer features. Some key techniques utilized in this sub-section include lemmatization, bag-of-words, TF-IDF, and named entity recognition. The methods described herein focus specifically on the free-text survey responses to the questions: "*What aspects of WHIMC contributed to finding the topic enjoyable, interesting, and/or easy to learn?"* and *"What aspects of WHIMC contributed to finding the topic uninteresting and/or challenging to learn?"*

3.6.1 Bag-of-words

Bag-of-words model is a straightforward and predominantly used method to extract features for machine learning models. This technique is easily adjustable and can be used in a lot of ways to extract features from texts. In this model, the collection of each text data instance will be represented as the bag (multiset) of its words. The grammar and word order will no longer be considered, but the frequency of the words will still be recorded [23]. Specifically, a bag-of-words is a representation of textual data that describes the occurrence of words within a dataset. It only has two components: the dictionary of all the unique words existing in the dataset which will be transformed into a column, and the frequency of each of those words row-wise. For the python implementation of Bag-of-words, a vectorizer function from the scikit-learn library [26] was utilized.

3.6.2 TF-IDF

TF-IDF, short for Term Frequency-Inverse Document Frequency, is another method used for extracting features from textual data. It consists of two components: TF and IDF [25]. Term Frequency (TF) measures how frequently a word appears in a document relative to the total number of words in that document. In the context of this study, a document corresponds to a survey response. The TF value is calculated by dividing the number of times a particular word appears in a document by the total number of words in that document which implies that the TF value increases as a word occurs more within a document. Conversely, Inverse Document Frequency (IDF) computes the importance of rare words in the dataset. This value is calculated for each word by taking the logarithm of the ratio between the total number of documents and the number of documents containing that word. This implies that scarcer words have higher IDF values. The final TF-IDF value is obtained by multiplying TF by IDF. The computational process for TF-IDF was accomplished in Python using a vectorizer function from the scikit-learn library [26].

3.6.3 Named Entity Recognition

Named Entity Recognition (NER) is the method of identifying nouns and proper nouns in a given text and categorizing them into predefined categories such as names of people, organizations, locations, dates, percentages, and more [19]. These categories function as additional features in the dataset, with the feature values denoting the frequency of each category within every survey response. Implementing NER in Python was achieved using the spaCy library [29].

3.7 Feature Selection

The researcher utilized Recursive Feature Selection (RFE) to automatically select a subset of the most relevant features from a given dataset. RFE works by recursively fitting the model to the data with different subsets of features and ranking them based on their importance, typically through a model-specific metric like coefficients in a logistic regression model or feature importances in a tree-based model [28].

3.8 Model Training and Evaluation

To determine which classifier should be used to predict a student's likelihood of re-engaging, the researcher applied five different classification models to the resulting dataset from the previous steps. These models include Logistic Regression, Random Forest,

Naïve Bayes, Support Vector Machine (SVM), and Multilayer Perceptron (MLP) [7]. Each of these models employs its unique approach to analyze the data and generate predictions for the target variable based on the given input variables. The performance of these models was then compared in terms of Sensitivity, Specificity, and ROC-AUC. Sensitivity measures the model's ability to correctly identify positive instances among all actual positives in the dataset, while Specificity assesses the model's ability to correctly identify negative instances among all actual negatives. The ROC curve visually represents a classifier's performance across different sensitivity and specificity thresholds. While it doesn't provide a single performance measure, the Area Under the Curve (AUC) condenses overall performance into a single metric. AUC values range from 0.5 to 1.0, with higher values indicating better performance [7, 17, 18].

4 RESULTS AND DISCUSSION

After completing the initial steps of the machine learning pipeline, including Data Collection, Data Cleaning, and Feature Engineering, a corpus of features ready to be trained by the selected models was produced. The dataset comprises 116 rows representing students and 176 features derived from both the Survey data (excluding free text) and in-game data. Additionally, 115 features originated from the free-text survey column, specifically from responses to the question "What aspects of WHIMC contributed to finding the topic enjoyable, interesting, and/or easy to learn?". The combination of these features then underwent Recursive Feature Elimination (RFE) before each round of modeling experiments, as detailed in the subsequent sections of this chapter.

The researcher aimed to predict if the student will re-engage with WHIMC or not. The target variable is binary and is defined as follows:

Table 1. Target Variable Definition

Class	Definition
1	The participant played WHIMC again the week after the 2-day training period.
0	The participant did not play WHIMC again after the 2-day training period.

In creating the models to predict student re-engagement, the researcher decided on specific sets of features to be used, as outlined below:

Table 2. Sets of Features Used in Prediction Models

F1 – Survey data (without the free text data) and in-game data (positions, observations, and science tools usage).
F2 – Survey data, features from the free text data, and in-game data (positions, observations, and science tools usage).
F3 – Features from the free text data only.

For each set of features above, 5 types of machine learning classifiers were trained on the data of 69 students, then tested on 47 hold out instances. Specifically, the machine learning classifiers used in this study were: (1) Logistic Regression, (2) Random Forest, (3) Naive Bayes, (4) Support Vector Machine (SVM), and (5) Multilayer Perceptron (MLP). To ensure optimal model performance across both classes, the metrics of Sensitivity, Specificity, and ROC-AUC were employed for evaluation.

The modelling experiments involved applying the chosen classifiers to each of the set of features as enumerated above. In total there were 15 classifiers made and assessed for this section. The results below show the metrics of the top 5 models ordered by ROC-AUC.

Table 3. Performance of Top 5 Classifiers

Model	Features	ROC-AUC
MLP	F1	.81
Logistic Regression	F1	.76
Naïve Bayes	F2	.71
SVM	F1	.68
Naïve Bayes	F3	.66

The most effective model for classifying student re-engagement in WHIMC was identified as the MLP classifier, employing survey data (without the free text data) and in-game data (positions, observations, and science tools usage) as features (F1). MLP is a type of artificial neural network capable of learning complex non-linear relationships in the data [32]. This is particularly useful when dealing with intricate patterns and interactions among features, which may be present in the dataset related to student re-engagement. Notably, for the best performing models, most of the features that were utilized are F1, alluding to its importance in predicting the outcome.

The confusion matrix below provides a detailed breakdown of the classification outcomes for the best MLP classifier in predicting student re-engagement:

Table 4. Confusion Matrix of the top performing MLP Classifier

		Predicted Label	
		0	1
Actual Label	0	11	3
	1	10	23

This model exhibited exemplary performance with an ROC-AUC of 0.81, sensitivity of 70%, and specificity of 79%. The choice of the MLP classifier was substantiated by its performance in discerning both positive and negative instances, achieving a balance between identifying students who would re-engage (sensitivity) and those who would not (specificity).

In particular, the emphasis on Specificity is crucial in this study, as this indicates that the model could be beneficial in providing early

intervention to students who are less likely to re-engage. Achieving a higher Specificity ensures a better ability to correctly identify true negatives, i.e., students who will not re-engage with WHIMC. This holds significance for facilitators, enabling them to target interventions toward students at a higher risk of disengagement. The consideration of Specificity in the model's performance underscores its practical application in education, facilitating proactive measures for students requiring additional support or encouragement. For instance, akin to the study by Hussain et al. [13], the model could be integrated into a system or deployed as a tool, allowing facilitators to monitor student engagement and offer timely interventions.

On the other hand, the results of the modeling experiments show that the addition of features extracted from the free text data for feature sets F2 and F3 did not improve the model performance. This could indicate that the quality and relevance of the textual features may not align with the patterns in the dataset related to student re-engagement. As a recommendation, exploring different techniques, such as more advanced NLP methods, could enhance the effectiveness of utilizing textual features. Additionally, incorporating more data, such as information from other schools implementing the WHIMC module, could enhance the model's ability to generalize and improve overall performance.

5 CONCLUSION

Re-engagement is defined as the learner's voluntary re-participation in an educational intervention. It is important as it is an indicator of intrinsic motivation which leads students to demonstrate enhanced learning effectiveness [8, 9, 38]. In this study, the specific challenge lies in developing a classifier that effectively gauges and distinguishes instances of student re-engagement by utilizing the data from a GBL called What-if Hypothetical Implementations in Minecraft (WHIMC). The researcher addressed this problem by investigating the following: **RQ 1:** How can we develop a machine learning pipeline to effectively predict a student's likelihood of re-engaging with WHIMC? **RQ 2:** How effectively does the best-performing machine learning model predict re-engagement?

In order to approach RQ1, the researcher provided a comprehensive methodology that details the approach and techniques to develop a machine learning classifier for student re-engagement. The data used for the study came from the module implementation of WHIMC in a middle school in the Philippines composed of a batch of Grade 8 students. Extracted data included in-game features, student demographics, and survey responses. To refine the data, the researcher corrected data types, handled missing values, standardized scales, and cleaned textual data. Feature engineering involved categorical data encoding, aggregation, and textual feature engineering using NLP methods such as bag-of-words, Named Entity Recognition (NER), and TF-IDF. Recursive Feature Elimination (RFE) was then applied to reduce dimensionality, preparing the data for model training. Various machine learning classifiers, including Logistic Regression, Random Forest, Naïve Bayes, Support Vector Machine (SVM), and Multilayer Perceptron (MLP) algorithms, were evaluated using ROC-AUC, Sensitivity, and Specificity.

In addressing RQ2, the researcher aimed to assess the efficacy of the models in predicting student re-engagement. This involved analyzing the relevant metrics from the modeling experiments to identify the best performing model. The researcher carefully selected specific sets of features for utilization. These feature sets, denoted as F1, F2, and F3, encompass different combinations of survey data, free text data, and in-game data, each tailored to capture distinct aspects. For each feature set, five types of machine learning classifiers were trained using the data from 69 students and subsequently tested them on 47 holdout instances to assess their predictive performance. The most effective model for classifying student re-engagement in WHIMC was identified as the MLP classifier, employing survey data (without the free text data) and in-game data (positions, observations, and science tools usage) as features (F1). This model exhibited exemplary performance with an ROC-AUC of 0.81, sensitivity of 70%, and specificity of 79%. The metrics of the resulting model highlight it has good performance in identifying true negatives, i.e., students who will not re-engage with WHIMC. This is important as this enables targeted interventions toward students at a higher risk of disengagement.

For future improvements on the study, the recommendation is to explore advanced NLP methods to enhance the effectiveness of using textual features. Additionally, incorporating more data, especially from other schools implementing the WHIMC module, could improve the model's generalization and overall performance.

REFERENCES

[1] Alonso-Fernandez, C., Calvo-Morata, A., Freire, M., Martinez-Ortiz, I. and Fernandez-Majón, B. 2019. Applications of data science to game learning analytics data: a systematic literature review. Computer Education. 141, 103612–103619.

[2] Astin, A.W. 1984. Student involvement: A developmental theory for higher education. Journal of College Student Development. 25, 297–308.

[3] Ayouni, S., Hajjej, F., Maddeh, M. and Al-Otaibi, S. 2021. A new ML-based approach to enhance student engagement in online environment. PLoS ONE. 16, (11) Nov. 2021, e0258788. DOI:https://doi.org/10.1371/journal.pone.0258788.

[4] Barak, M. 2018. Are digital natives open to change? Examining flexible thinking and resistance to change. Computers & Education. 121, 115–123.

[5] Bertozi, E. 2014. Using Games to Teach, Practice and Encourage Interest in STEM Subjects. Learning, Education and Games. 23–36.

[6] Bond, M. and Bedenlier, S. 2019. Facilitating student engagement through educational technology: Towards a conceptual framework. Journal of Interactive Media in Education. 1, (11) 2019, 1–14.

[7] Bruce, P. and Bruce, A. 2017. Practical Statistics for Data Scientists: 50 Essential Concepts. O'Reilly Media, Inc.

[8] Cairns, P. 2016. Engagement in digital games. In Why engagement matters,. 81–104.

[9] Casano, J.D.L., Fuentes, M. and Rodrigo, M.M.T. 2023. Quantifying Re-engagement in Minecraft. Artificial Intelligence in Education. 1831.

[10] Chu, X., Ilyas, I.F., Krishnan, S. and Wang, J. 2016. Data Cleaning: Overview and Emerging Challenges. Proceedings of the 2016 International Conference on Management of Data (New York, NY, USA, Jun. 2016), 2201–2206.

[11] Domina, T., Renzulli, L., Murray, B., Garza, A.N. and Perez, L. 2021. Remote or removed: Predicting successful engagement with online learning during COVID-19. Socius: Sociological Research for a Dynamic World. 7.

[12] Henderson, M., Selwyn, N. and Aston, R. 2017. What works and why? Student perceptions of 'useful' digital technology in university teaching and learning. Studies in Higher Education. 42, (8) 2017, 1567–1579.

[13] Hussain, M., Zhu, W., Zhang, W. and Abidi, S.M. 2018. Student engagement predictions in an e-learning system and their impact on student course assessment scores. Computational Intelligence and Neuroscience. 1–21.

[14] Khyani, D. and B S, S. 2021. An Interpretation of Lemmatization and Stemming in Natural Language Processing. Shanghai Ligong Daxue Xuebao/Journal of University of Shanghai for Science and Technology. 22, 350–357.

[15] Kuh, G.D. 2001. Assessing what really matters to student learning: Inside the national survey of student engagement. Change. 33, (3) 2001, 10–17.

[16] Kuh, G.D., Cruce, T.M., Shoup, R., Kinzie, J. and Gonyea, R.M. 2008. Unmasking the effects of student engagement on first-year college grades and persistence. Journal of Higher Education. 79, (5) 2008, 540–563.

[17] Melo, F. 2013. Area under the ROC Curve. Encyclopedia of Systems Biology. W. Dubitzky, O. Wolkenhauer, K.-H. Cho, and H. Yokota, eds. Springer. 38–39.

[18] Melo, F. 2013. Receiver Operating Characteristic (ROC) Curve. Encyclopedia of Systems Biology. W. Dubitzky, O. Wolkenhauer, K.-H. Cho, and H. Yokota, eds. Springer. 1818–1823.

[19] Mohit, B. 2014. Named Entity Recognition. Natural Language Processing of Semitic Languages. I. Zitouni, ed. Springer. 221–245.

[20] NLTK :: Natural Language Toolkit. Retrieved November 10, 2023 from https://www.nltk.org/.

[21] O'Brien, H. and Toms, E.G. 2008. What is user engagement? A conceptual framework for defining user engagement with technology. Journal of the American Society for Information Science and Technology. 59, (6) 2008, 938–955.

[22] pandas documentation — pandas 2.1.1 documentation. Retrieved October 26, 2023 from https://pandas.pydata.org/docs/.

[23] Qader, W., M. Ameen, M. and Ahmed, B. 2019. An Overview of Bag of Words;Importance, Implementation, Applications, and Challenges. (Jun. 2019), 200–204.

[24] Ross, A. and Willson, V.L. 2017. Descriptive Statistics. Basic and Advanced Statistical Tests: Writing Results Sections and Creating Tables and Figures. A. Ross and V.L. Willson, eds. SensePublishers. 3–7.

[25] Sammut, C. and Webb, G.I. eds. 2010. TF–IDF. Encyclopedia of Machine Learning. Springer US. 986–987.

[26] scikit-learn: machine learning in Python — scikit-learn 1.3.2 documentation. Retrieved November 10, 2023 from https://scikit-learn.org/stable/.

[27] Selwyn, N. 2016. Exploring university students' negative engagements with digital technology. Teaching in Higher Education. 21, (8) 2016, 1006–1021.

[28] sklearn.feature_selection.RFE. Retrieved November 6, 2023 from https://scikit-learn/stable/modules/generated/sklearn.feature_selection.RFE.html.

[29] spaCy · Industrial-strength Natural Language Processing in Python. Retrieved November 10, 2023 from https://spacy.io/.

[30] Tablatin, C.L.S., Casano, J.D.L. and Rodrigo, M.M.T. 2023. Using Minecraft to cultivate student interest in STEM. Frontiers in Education. 8.

[31] Taub, M., Sawyer, R., Smith, A., Rowe, J., Azevedo, R. and Lester, J. 2020. The agency effect: The impact of student agency on learning, emotions, and problem-solving behaviors in a game-based learning environment. Computers & Education. 147.

[32] Taud, H. and Mas, J.F. 2018. Multilayer Perceptron (MLP). Geomatic Approaches for Modeling Land Change Scenarios. M.T. Camacho Olmedo, M. Paegelow, J.-F. Mas, and F. Escobar, eds. Springer International Publishing. 451–455.

[33] Tromba, P. 2013. Build Engagement and Knowledge One Block at a Time with Minecraft. Learning & Leading with Technology. 40, (8) 2013, 20–23.

[34] Verspoor, K. and Cohen, K.B. 2013. Natural Language Processing. Encyclopedia of Systems Biology. W. Dubitzky, O. Wolkenhauer, K.-H. Cho, and H. Yokota, eds. Springer. 1495–1498.

[35] What-if hypothetical implementations in minecraft. from https://whimcproject.web.illinois.edu/.

[36] WHIMC Teacher Guides. Retrieved October 24, 2023 from https://whimcproject.web.illinois.edu/education-research/teacherguide/.

[37] Yi, S. 2021. The impacts of a science-based videogame intervention on interest in STEM for adolescent learners.

[38] Zaccone, M.C. and Pedrini, M. 2019. The effects of intrinsic and extrinsic motivation on students learning effectiveness. Exploring the moderating role of gender. International Journal of Educational Management. 33, (6) 2019, 1381–1394.