Forecasting the performance of early childhood education childrens through a game-based learning approach

Gabriel Candido da Silva, Universidade Federal Rural de Pernambuco, gabcandidods@gmail.com, 0000-0002-7470-8613
Rodrigo Lins Rodrigues, Universidade Federal Rural de Pernambuco, rodrigomuribec@gmail.com, 0000-0002-3598-5204
Américo Nobre Amorim, Escribo – Inovação para o Aprendizado, americo@escribo.com, 0000-0002-7834-2057
Amadeu Sá de Campos Filho, Universidade Federal de Pernambuco, amadeu.campos@gmail.com, 0000-0002-8660-554X

Abstract: Currently, research that seeks to evaluate the learning acquired by players from a Serious Game has been adopting measures to demonstrate evidence collected in real-time. However, few studies seek to carry out this type of evaluation and techniques in Serious Games for early childhood education. That said, this study sought to apply a Game Learning Analytics process to assess at what level it is possible to predict the learning effect of 331 preschool childrens, based only on their interaction characteristics with 20 games that work on reading and writing skills, also showing which were the most effective interaction characteristics and classification techniques for this task. We found that errors in the games are the most relevant characteristic, and the Random Forest classifier is the most suitable for this experiment, rating a Precision of 82%.

Keywords: Early years education, Serious games, Game learning analytics, Classification

Previsão do desempenho de crianças da educação infantil por meio de uma abordagem de aprendizagem baseada em jogos

Resumo: Atualmente, pesquisas que buscam avaliar o aprendizado adquirido pelos jogadores a partir de um Serious Game vêm adotando medidas para demonstrar evidências coletadas em tempo real. Contudo, poucos estudos buscam realizar este tipo de avaliação e técnicas em Jogos Sérios para a educação infantil. Dito isto, este estudo procurou aplicar um processo de Game Learning Analytics para avaliar em que nível é possível prever o efeito de aprendizagem de 331 crianças do ensino pré-escolar, com base apenas nas suas características de interação com 20 jogos que trabalham as competências de leitura e escrita, mostrando também quais foram as características de interação e técnicas de previsão mais eficazes para esta tarefa. Descobrimos que os erros nos jogos representam a característica mais relevante, e o classificador Random Forest é o mais adequado para este experimento, com uma precisão de 82%.

Palavras-chave: Educação infantil, Jogos sérios, Análise de aprendizagem em jogos, Classificação

1. INTRODUCTION

Digital games can be understood as a form of interactive and engaging media, which requires attention and creativity in complex processes execution, developing several positive effects on the player, such as improvements in perception, attention, memory, and decision-making (EICHENBAUM et al., 2014). In addition, the constant growth of research in digital games has shown the effectiveness of its application in learning environments; and that one can increasingly understand how to use their characteristics to contribute to education.

Digital games aiming at educational purposes can fit into the definition of Serious Games (SG), which are games created with purposes that go beyond entertainment, where players learn and develop skills by overcoming challenges during the act of playing (ZHONGGEN, 2019). The use of Serious Games as tools to aid learning can impact the entire children's perspective, working on essential skills in contemporary society, such as critical thinking, problem-solving, and decision making.

In general, games naturally constitute a very interactive environment, and this combines with the application of data science techniques because environments such as those of a game can generate (and then captured) different types of data (ALONSO-FERNÁNDEZ et al., 2019). Due to the enormous capacity of the SG to generate information, there is a specific area of data science for performing analysis of learning in games called Game Learning Analytics (GLA), which is the process of collecting, analyzing, and understanding data in Serious Games.

The systematic review of the literature carried out by Alonso-Fernández et al. (2019), which sought to reveal an overview of data science applications in GLA, demonstrates that the Assessment category, which encompasses the Assessment of Acquired Learning and Performance Prediction, is the most used study category in the area, being present in 32 of the 87 studies analyzed in the survey. In addition, it is also possible to identify a gap concerning the educational level of the participants in these interventions, demonstrating that none of the 87 studies analyzed sought to cover early childhood education.

Therefore, this study aims to apply a Game Learning Analytics process to fill this gap. It seeks to evaluate at what level it is possible to predict the learning effect on childrens from early childhood education using only their interaction characteristics with games. And to identify which of these interaction characteristics (as well as the techniques used for classification) are more effective for an adequate application of the process, thus allowing teachers to identify which childrens need pedagogical assistance during the learning process. For this, the following research questions were raised:

- Q1: At what level is it possible to classify the performance of childrens from early childhood education through the characteristics of their interaction with digital games?
- Q2: Which game interaction characteristics are most relevant, and which classification techniques are most effective?

2. RELATED WORKS

In studies that seek to carry out assessments of the learning acquired by players from a Serious Game, it is common to apply external questionnaires to measure knowledge, but measures are currently adopted to demonstrate evidence collected in real-time, while childrens are still playing (ALONSO-FERNÁNDEZ et al., 2023). The

study of Chen et al. (2020) demonstrates that some studies seek to use sequences of children actions in the game, together with other characteristics, to predict post-test performance using Machine Learning and Deep Learning techniques.

In the work of Satu et al. (2021), the authors developed a game application called COVID-Hero to raise awareness among children about COVID-19, and also conducted questionnaires to collect opinions and identify the most important characteristics. For this, the data from these questionnaires were analyzed using 5 different algorithms and among the results obtained, the use of XGBoost stood out, which was estimated as the best for the experiment with the maximum value of R-Squared (0.950) and the lower residuals. The second best was Random Forest with a score of R-Squared (0.890).

The Random Forest model was also used by Dapogny et al. (2018) to classify facial expressions that are presented in JEMImE, a serious game solution that aims to teach children with autism spectrum disorder (ASD) how to produce facial expressions. For this, in the study, three databases of videos portraying four classes of children's PEs were labeled: neutral, happiness, anger and sadness.

In Chen et al. (2020), the authors aimed to identify behavioral characteristics as indicators of childrens' game activities. For this, they used the SVM, which the authors highlighted as an advantageous choice given the characteristics of their dataset. As a result, the SVM presented the five most important characteristics with an AUC rate of 67.32%, these characteristics are related to the number of clicks and attempts. However, the authors cite as a limitation that the DGBA used has a limited number of childrens for predictive modeling.

The research of Shin et al. (2020) apply the classification techniques in a tablet game designed for children from Pre-K to 2nd grade. Their work shows that with in-game motion detection, it is possible to distinguish and predict childrens with "solver" behavior from the activities of those "guessers". However, it is noticeable that the accuracy values obtained need to be improved so that some automation of the method can be carried out on a large scale.

Another characteristic present in research that seeks to assess and predict the knowledge acquired by childrens during the act of playing is the lack of longitudinal applications of games, such as the case study carried out by Alonso-Fernández et al. (2020), which sought to measure the learning acquired by childrens based on interactions with the game. However, the entire process of applying the pre-post test method took place in a 50-minute interval, where childrens had 10 minutes for each test and 30 minutes to use the game.

Given this scenario, this study seeks to fulfill the gaps left by previous studies, which demonstrate a lack in the application of experiments with early childhood education, in addition to other application problems that can make the results of the studies biased, such as the low number of participating schools and the short period in which the experiment is applied.

3. METHOD

3.1 Context

The research proposed in this study was developed from data collected from the Escribo Play application, an educational game platform for mobile devices, with more than 400 games that contribute to the development of language, math, and science skills, and built to be used by early childhood teachers and childrens.

The data used in this study were collected in a secondary research, extracted from the execution of an experiment that applied a specific set of twenty games from the Escribo Play, built to develop skills related to phonological awareness (PA). The study conducted by Amorim et al. (2020) addresses this experiment, which sought to evaluate the effectiveness of using these games from a pre-test and post-test experiment to measure the learning gain of 749 4-year-old preschool childrens from 15 private schools located in 5 different cities from a metropolitan region in Northeast Brazil.

The 749 childrens who participated in the experiment were divided into two groups: an experimental group, which was the group that received the intervention and played the 20 selected games, and a control group, who received only the usual classes taught by their teachers. In addition to the experiment application, data on the experimental group (331 childrens) game interaction were collected, such as the number of views, hits, and errors in each game.

The 20 games worked on in this intervention can be divided into 4 categories: 1) Games that work on skills related to syllables (n = 4), such as joining, separating, adding and inverting syllables; 2) Games that work on skills related to Rhymes and Alliteration (n = 4); 3) Games that work skills related to Phonemic Awareness (n = 6) and; 4) Games that work on Reading and Writing skills (n = 6).

3.2 Game learning analytics approach

This study seeks to address a proposal for evaluating the learning gain during the process of using the games, where the characteristics of interaction with these games will be explored to build a classification model of the performance of the childrens who participated in the experiment. Figure 1 presents a flowchart of the intervention process and interaction data extraction, as well as the Game Learning Analytics proposal to model this data.

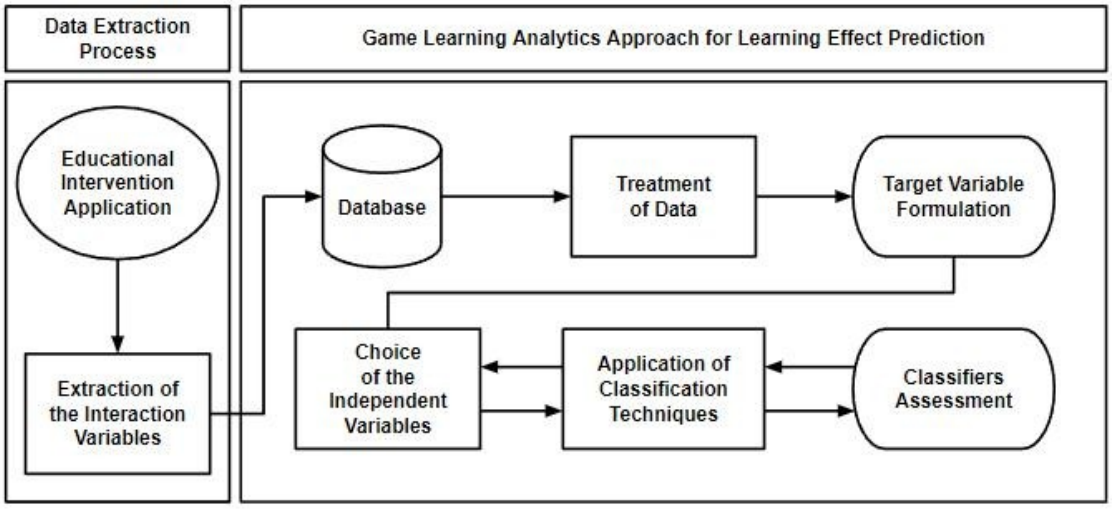

Figure 1 - Flowchart of the method proposed in the study.

As shown in the flowchart, the process begins with the data extraction stage, which occurs during the experiment carried out by Amorim et al. (2020). After accessing the database, a process of data analysis and cleaning occurred to identify and remove childrens who have NA values in any of their variables so that they do not interfere with the experiment. Table 1 shows the description of the collected variables that will be used in this research.

Table 1 - Description of the variables collected.

Variable	Description
Reading_Pre	Pre-test reading score
Reading_Post	Post-test reading score
Writing_Pre	Pre-test writing score
Writing_Post	Post-test writing score
Activity*_V	Number of views in the game
Activity*_R	Number of hits in the game
Activity*_W	Number of errors in the game
Gain_R	Value obtained from: (Reading Post - Reading Pre)
Gain_W	Value obtained from: (Writing Post - Writing Pre)

All variables described in Table 1 are of numeric type, and the term "Activity*" refers to the game in question, among a universe of 20 different games, that is, for each game, there is a number of hits, errors, and views. From the use of these variables, a process of analysis and treatment of the data will be carried out to remove subjects with NA values from the table because the variable to be used for the classification (the gain) is dependent on the reading and writing pre-test and post-test, and a value of NA in these variables means the children did not participate in the tests.

After this step, we started to build the classification model. Our goal with the classification model is to be able to differentiate childrens who can adequately receive the effects expected from the intervention and have a positive learning effect from those who cannot. We do this to quickly identify those childrens who are having difficulties during the intervention in order to apply the necessary pedagogical measures.

To carry out the application, we used Classification techniques, which consists of a set of machine learning methods to compute the probability of an individual belonging to a certain class based on one or multiple predictor characteristics (KASSAMBARA, 2017). We chose to use the classification because we believe it is more appropriate, given the context of our research, to understand which childrens are obtaining a positive learning effect or not.

To proceed with the classification model, a categorical target variable was constructed. Given the purpose of our study, we used the learning gain related to the value of the post-test minus the pre-test obtained by each children to categorize the childrens into two classes that represent whether there was a learning effect or not. For this, it was considered that there was a learning effect in cases where the reading or writing gain was higher than 0; and that there was no learning effect when the gain was less than or equal to 0.

The choice of independent variables was made by the combinations of hits, errors, and views of the twenty games, because these are the variables related to the use of games and consequently the ones who better demonstrate how each children interacted with the application.

As there is a target variable for the gain in writing and another target variable for the gain in reading, the entire process of combinations of variables for the application of classification techniques involved two approaches: in the first approach, the possible combinations of the independent variables were performed targeting the effect on reading; and, later, in the second approach, they targeted the effect in writing.

To apply the classification models, the database containing all research subjects was divided into training and test data, 70% for training and 30% for testing and the classification algorithms were applied at the end of the model training process. Given this, and the examples of studies in the context of education evidenced in the Related Works section, we chose 4 classification algorithms that are commonly used: Generalized Linear Model (GLM), Support Vector Machine (SVM), Random Forest (RF) and XGBoost (XGB).

4. RESULTS

As described previously, four different classification algorithms were used. Each of these ran for seven different combinations of variables. In the first test round, the target variable was the reading effect, and in the second test round, the target variable was the writing effect, which totaled 56 different outputs. Among these outputs are the values of the metrics of the classification performed. The metrics of Accuracy, Kappa, Recall, and Precision were extracted from these values.

After running the all-possibility test, the values of the rating metrics mentioned above were tabulated. It was possible to notice a lower-than-expected performance of the algorithms when the target variable was modeled as the effect on childrens' writing skills. Thus, it was decided to go with only the classification data of the effect on reading ability, which has 36% of the sample classified as 0 (had no positive learning effect) and 64% as 1 (had positive learning effect), which demonstrates a reasonable balance of the class. Table 2 shows the ten best results obtained, ordered according to Accuracy.

Table 2 - Best results obtained by classification techniques.

Model	Combination	Accuracy	Kappa	Recall	Precision
XGB	V + W	**0.75**	**0.41**	**0.54**	0.67
XGB	V+R+W	**0.75**	**0.38**	**0.42**	0.73
RF	V + W	**0.75**	**0.35**	0.35	**0.82**
RF	V+R+W	**0.75**	**0.35**	0.35	**0.82**
XGB	R	0.73	**0.35**	**0.42**	0.69
RF	R + W	0.73	0.31	0.31	**0.80**
RF	W	0.72	0.30	0.35	0.69
GLM	W	0.72	0.29	0.31	0.73
RF	V	0.72	0.27	0.27	0.78
SVM	W	0.71	0.19	0.15	1.00

The three best values obtained for each metric are highlighted in bold. The table is ordered according to Accuracy, which reports in percentage how many ratings, negative and positive, were correctly predicted. In terms of Accuracy, it is possible to notice a highlight of the Random Forest and XGBoost classifiers, notably when used with the the variables "V + R + W" and "V + W" respectively. For Kappa, which shows the individuals degree of agreement, the best results regarding Accuracy are also repeated, which only reinforces the results obtained since, as demonstrated in the study

by Classe & Castro (2020), Kappa values between 0.21 and 0.40 are considered medium or fair and between 0.41 and 0.60, moderate.

The table also shows the Recall metric, which demonstrates the percentage of accurateness of the forecast considering all positive classes, which includes true positives and false negatives, and Precision, which considering all predicted classes as positive, that is, the true and false positives, shows the percentage of how many are actually positive. For these two metrics, there is once again a good performance of the best combinations for Accuracy and Kappa, with emphasis on combinations of variables made with the XGBoost classifier, which obtained the best values for Recall, and the Random Forest classifier, which obtained the best values in the Precision metric.

The ROC curve was generated after surveying for the best combinations, and the calculation of the Area Under the Curve (AUC) was performed based on the best combination of variables for each classifier, considering the Accuracy. The best combinations were: 1) Random Forest with "V + W" combination (Accuracy = 0.75); 2) XGBoost with "V + W" combination (Accuracy = 0.75); 3) GLM with "W" combination (Accuracy = 0.72) and; 4) SVM with "W" combination (Accuracy = 0.71).

Figure 2 shows the ROC curve of the four classifiers from the combination of variables with the best accuracy in each one of them.

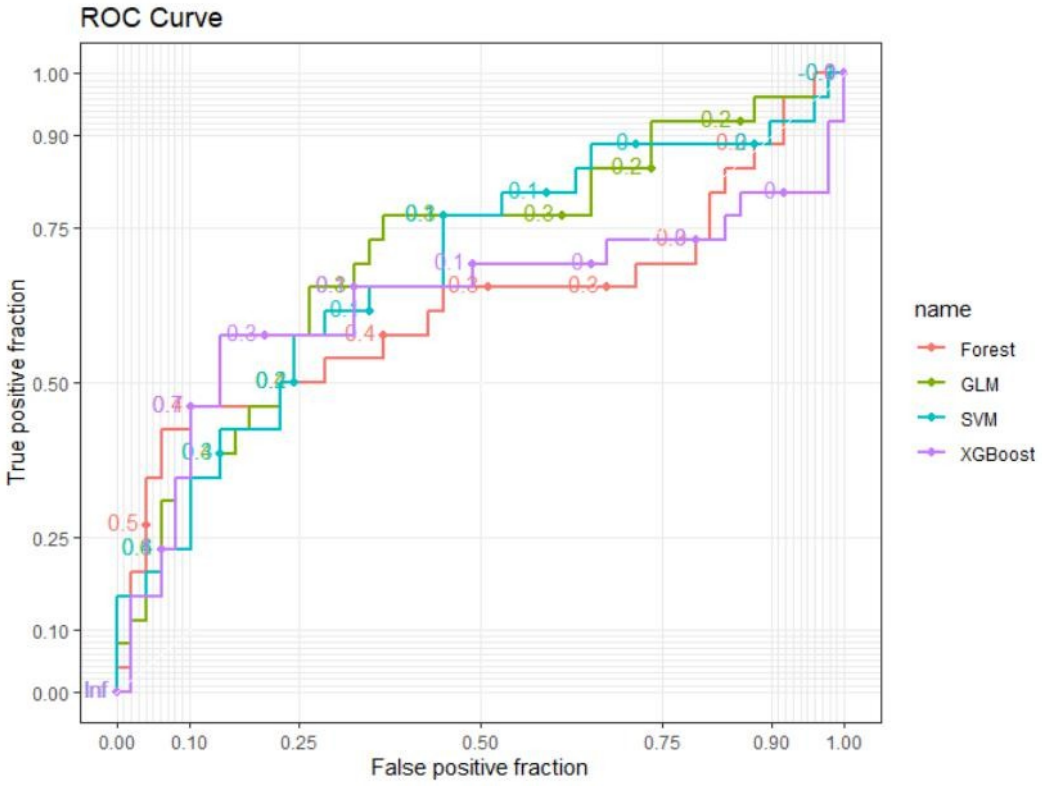

Figure 2 - ROC Curve of the four classifiers used.

To better understand what this visualization means, the AUC values were generated. It is a graphical representation of the ROC Curve, where the higher the AUC value, the better this classifier performed in the classification. The values obtained for each classifier were: RF = 0.62; GLM = 0.70; SVM = 0.68; XGBoost = 0.63.

5. DISCUSSION

This study sought to carry out a Game Learning Analytics process to verify at what level it is possible to predict the learning effect and which are the interaction characteristics and classification algorithms that best perform this task. Thus, this section seeks to answer the research questions raised in Section 1, based on an educational discussion of the results achieved.

To answer Q1, we performed a classification process, which tested the seven possible combinations of variables and four different classification models. Considering the best values shown in Table 2, it was obtained: Accuracy = 0.7467; Recall = 0.5385; Precision = 0.8182. These values are within the expected range when compared to the study of Shin et al. (2020), whose sample has characteristics that are similar to this study, having obtained as the best values: Accuracy = 0.778; Recall = 0.781; Precision = 0.777. We also present in figure 3 the respective confusion matrices generated for the best combination of each algorithm.

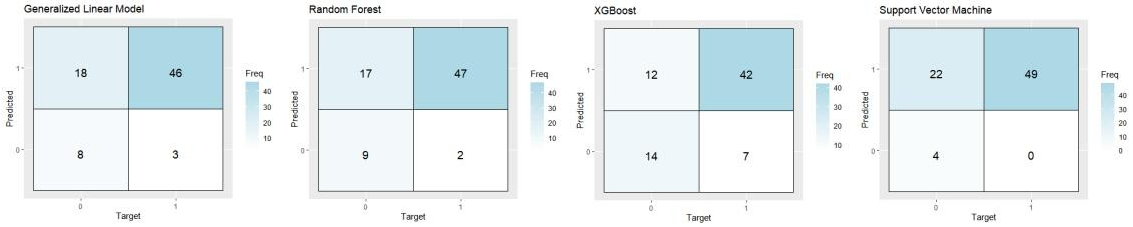

Figure 3 - Confusion matrices of the best combinations of variables for each algorithm

When analyzing the results of studies such as those by Juric et al. (2021), which achieved Accuracy values above 0.90, or by Alonso-Fernández et al. (2020), with values above 0.90 in Recall and Precision, it is noticeable that the values of this study are below those of related works when compared to the values of research in which the target audience was not early childhood education childrens.

That said, it is understood that the robustness of the experiment, which occurred with the data collected from ten weeks and had the participation of 331 childrens from 15 different schools, and the fact that the childrens belong to kindergarten, a group rarely addressed in related research, make the results obtained quite promising.

In addition to Accuracy reaching a value within a reasonable and expected range for this context, a Precision of 0.8182 (81%) was obtained, which points to it as a good predictor for correctly classifying childrens who obtained a positive learning effect. It means that it is unlikely that a children who has no learning effect will be classified incorrectly. However, Recall was low, reaching only 0.5385 (53%), which points to the predictor as relatively flawed to accurately classify childrens who did not get a learning effect.

Regarding Q2, whose objective is to identify which interaction characteristics and classification models were most effective for the process, a survey of the ten best combinations was carried out, based on the Accuracy metric, which indicates how many classes were correctly predicted.

In Table 2, it is possible to notice that the characteristic that is most repeated is that of errors in the games (“W”), being present in 8 of the ten best combinations raised. In addition, the error characteristics are also present in the best result of each classifier, which shows how important it is (considering this database) to use error variables in games, reinforcing the finding addressed in the research by Silva et al. (2022), who applied a cluster analysis on this same database and found that the characteristics of errors are the most determinant to generate groups with different behavioral profiles.

In addition, XGBoost and Random Forest are the classifiers with the best Accuracy, tied at 74%. Both also have good values in the Kappa metric, reaching above 0.30, which reinforces the results obtained. The main difference between the two techniques is in Recall, where XGboost performed better, and Precision, where the Random Forest classifier performed better.

The priority for this research context is to adopt a classifier in which the cost of false positives is high because, understanding that the objective is to identify those childrens who need help during the process of using the games, one wants to avoid as much as possible that childrens who do not have a learning effect are classified as having an effect. Therefore, the Random Forest classifier was chosen as the most suitable because it reached the highest Precision rate, 81%.

6. CONCLUSIONS

This study sought to assess at what level it is possible to predict the learning effect of preschool childrens participating in an experiment that applied a set of 20 games to develop reading and writing skills. In addition, it sought to identify which characteristics of interaction with these games and which classification models were more effective to obtain high rates in the metrics of Accuracy, Kappa, Precision, and Recall.

In order to answer Q1, the best values obtained in the metrics were compared to those of other similar studies to verify that the results shown in this study are within the expected for the context of early childhood education, also considering the robustness of the experiment application that served for data collection. Regarding Q2, comparing the results obtained by the different combinations of variables and classifiers that were performed in this experiment, it is possible to notice that the error variables in games are the most relevant characteristics and that the Random Forest classifier is the model best suited for this context.

Regarding the limitations of our study, the process performed and the results obtained refer to a single dataset that is composed of 331 4-year-old childrens, in which the available interaction data attributes were visualizations, successes and errors in 20 games that specifically aim to develop phonological awareness, which may imply in generalization difficulties in other learning domains, with children of different age groups and with different interaction characteristics captured by the game. Finally, the randomized controlled study in which data were collected included only private schools serving middle-class families. Future studies may perform similar analyzes and comparisons with other game-based interventions in other learning domains, or even when delivered to childrens in poverty.

From this study, it will be possible to implement the models developed in real-time to build tools and applications that help teachers identify childrens who need help even during the learning process. Thus, contribute to avoiding the use of evaluative instruments based on evidence and tests, which are expensive, and need time and qualified people to be applied.

REFERENCES

ALONSO-FERNÁNDEZ, C. et al. Applications of data science to game learning analytics data: A systematic literature review. **Computers & Education**, v. 141, p. 103612, 2019.

ALONSO-FERNÁNDEZ, C. et al. Evidence-based evaluation of a serious game to increase bullying awareness. **Interactive Learning Environments**, v. 31, n. 2, p. 644-654, 2023.

ALONSO-FERNÁNDEZ, C. et al. Predicting students' knowledge after playing a serious game based on learning analytics data: A case study. **Journal of Computer Assisted Learning**, v. 36, n. 3, p. 350-358, 2020.

AMORIM, A. N., JEON, L., ABEL, Y., FELISBERTO, E. F., BARBOSA, L. N. F., & DIAS, N. M. (2020). Using Escribo Play Video Games to Improve Phonological Awareness, Early Reading, and Writing in Preschool. Educational Researcher, 0013189X2090982. doi:10.3102/0013189x20909824

CHEN, F.; CUI, Y.; CHU, M.. Utilizing game analytics to inform and validate digital game-based assessment with evidence-centered game design: A case study. **International Journal of Artificial Intelligence in Education**, v. 30, p. 481-503, 2020.

CLASSE, T., & CASTRO, R. Técnicas e conceitos de business intelligence para avaliação em jogos educacionais: Um mapeamento sistemático da literatura, **Proceedings of SBGames 2020**, 2020

DAPOGNY, A. et al. JEMImE: a serious game to teach children with ASD how to adequately produce facial expressions. In: **13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)**. IEEE, 2018. p. 723-730.

EICHENBAUM, A.; BAVELIER, D.; GREEN, C. S. Video games: play that can do serious good. **American Journal of Play**, v. 7, n. 1, p. 50-72, 2014.

HAN, J.; PEI, J.; TONG, H. **Data mining: concepts and techniques**. Morgan kaufmann, 2022.

JURIC, P.; BAKARIC, M. B.; MATETIC, M. Detecting students gifted in mathematics with stream mining and concept drift based m-learning models integrating educational computer games. **International Journal of Emerging Technologies in Learning (iJET)**, v. 16, n. 12, p. 155-168, 2021.

KASSAMBARA, A. **Practical guide to cluster analysis in R: Unsupervised machine learning.** Sthda, 2017.

SATU, M. S. et al. COVID-Hero: machine learning based COVID-19 awareness enhancement mobile game for children. In: **International Conference on Applied Intelligence and Informatics.** Cham: Springer International Publishing, 2021. p. 321-335.

SHIN, H.; KIM, B.; GWEON, G. Guessing or Solving? Exploring the Use of Motion Features from Educational Game Logs. In: **Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems**. 2020. p. 1-8.

SILVA, G. C., RODRIGUES, R. L., AMORIM, A. N., MELLO, R. F., NETO, J. R. O. Game learning analytics can unpack Escribo play effects in preschool early reading and writing. Computers and Education Open, Volume 3, 2022, 100066, ISSN 2666-5573, 2022. oi: 10.1016/j.caeo.2021.100066

ZHONGGEN, Y. et al. A meta-analysis of use of serious games in education over a decade. **International Journal of Computer Games Technology**, v. 2019, 2019.