Education and Information Technologies (2024) 29:15793–15812
https://doi.org/10.1007/s10639-024-12500-w

A machine learning based model for student's dropout prediction in online training

Meriem Zerkouk[1] · **Miloud Mihoubi**[2] · **Belkacem Chikhaoui**[2] · **Shengrui Wang**[1]

Received: 11 July 2023 / Accepted: 18 January 2024 / Published online: 2 February 2024

Abstract

School dropout is a significant issue in distance learning, and early detection is crucial for addressing the problem. Our study aims to create a binary classification model that anticipates students' activity levels based on their current achievements and engagement on a Canadian Distance learning Platform. Predicting student dropout, a common classification problem in educational data analysis, is addressed by utilizing a comprehensive dataset that includes 49 features ranging from socio-demographic to behavioral data. This dataset provides a unique opportunity to analyze student interactions and success factors in a distance learning environment. We have developed a student profiling system and implemented a predictive approach using XGBoost, selecting the most important features for the prediction process. In this work, our methodology was developed in Python, using the widely used sci-kit-learn package. Alongside XGBoost, logistic regression was also employed as part of our combination of strategies to enhance the models predictive capabilities. Our work can accurately predict student dropout, achieving an accuracy rate of approximately 82% on unseen data from the next academic year.

✉ Meriem Zerkouk
meriem.zerkouk@usherbrooke.ca

Miloud Mihoubi
miloud.mihoubi@teluq.ca

Belkacem Chikhaoui
belkacem.chikhaoui@teluq.ca

Shengrui Wang
shengrui.Wang@usherbrooke.ca

[1] Department of Computer Science, University of Sherbrooke, Street, Sherbrooke 100190, Quebec, Canada

[2] Artificial Intelligence Institute, University of Téluq, 5800, rue Saint-Denis, bureau 1105, Montreal H2S 3L5, Quebec, Canada

Keywords Dropout school · Machine learning · Prediction · Sociodemographic data · Bihavioral data

1 Introduction

Student dropout is a global problem with negative consequences for individuals, families, and communities. Approximately 10,000 young people in Quebec, Canada, leave school each year without obtaining a first diploma (scolaire, 2023). The lifetime cost of a school dropout to society is 120 million per annual cohort. This has a significant economic impact, as school dropouts have difficulty finding employment and earn lower incomes. Therefore, it is essential to implement strategies to prevent dropouts and help them persevere in returning to school (Alam et al., 2023). Major factors in school dropout include internal factors such as learning difficulties, mental or behavioral problems, and external factors such as financial difficulties, housing problems, or insufficient family support (Shiao et al., 2023). Past learning systems used by our platform considered school absenteeism and poor academic performance as key indicators of school dropout. However, it is important to recognize that these indicators are not the only factors contributing to school dropout.

Distance learning (Oz et al., 2022) has become an important and suitable solution to help students who have dropped out of school resume their studies. It allows access to education for students who cannot attend traditional school for various reasons (distance, health, family situation, work, etc.). However, more than distance learning is needed to solve school dropouts definitively. Adapting strategies to accommodate distance learning is crucial to avoiding the risk of dropping out.

Distance learning systems offer an alternative way for students to complete their high school diploma through distance learning using an advanced online platform. It provides personalized support, resources, and services to help students overcome the challenges that caused them to drop out of traditional school, including support, encouragement, motivation, and building self-confidence to return to their studies. Our vision for preventing school dropout through a preventive approach is based on two main axes: integrating an intervenor to support students and identifying students at risk of dropping out based on our knowledge of risk factors or relevant attributes to drop out.

To prevent and effectively address school dropout, it is necessary to identify key dropout factors using artificial intelligence techniques. We can support students who need assistance by targeting modifiable causes and providing effective interventions. Our research offers a solution for preventing and combating school dropout by using data analysis and machine learning techniques to identify at-risk students. It allows us to implement personalized interventions to help them persevere in their studies. Moreover, by using data from various sources, we can evaluate our interventions' effectiveness and continuously improve student results.

Our contribution to addressing the problem of student dropout in distance learning consists of a three-step approach:

- **Identification of factors:** We begin by analyzing available educational data to identify the key factors contributing to student These factors may include demographic information, past academic performance, and engagement with learning materials.
- **Development of student profiles:** Based on the identified factors, we create distinct student profiles that can help in understanding the specific needs and challenges of different shapes of students at risk of dropping out.
- **Prediction:** We implement a prediction approach using a machine learning algorithm (XGBoost), considering the most important features identified in the previous steps. It allows us to accurately predict student dropout and implement personalized interventions to help them persevere in their studies.

This article is organized as follows: the first section describes related work on the topic of school dropout, the second section describes the methodology followed for data analysis, the third section outlines the development of artificial intelligence algorithms, and finally, the fourth section presents the obtained results.

2 Related work

Online education, also known as distance learning, offers a flexible and accessible alternative to traditional classroom learning. However, there may be a high dropout rate in online learning, where students drop out of the online courses they have started. Predicting dropout involves several research areas, such as education and machine learning. The specific objective of this study is to analyze dropouts in remote learning environments by modeling the profile and behavior of students when they interact with online learning platforms. To demonstrate the contribution of our work in this field, we review a set of studies, analyzing them based on three main aspects: Learning platform and collected data, student modeling (profile and behavior), and prediction method. At the end of our study, we can gain a deep understanding of the factors contributing to dropout, develop more effective prediction algorithms to help at-risk students and implement more relevant intervention strategies. The different research works were analyzed regarding distance learning platforms, student profile modeling, and machine learning techniques used to identify school dropouts.

- For the objective of predicting student dropout from online courses (Prenkaj et al., 2020), we did a thorough analysis of machine learning methods. The survey covers a variety of ways, including deep learning, support vector machines, decision trees, random forests, and logistic regression. An early dropout prediction model for programming courses supported by online judges was put forth by Bonifro et al. (2020). Their approach incorporates elements taken from information on student behavior and data from programming exercises. A student dropout prediction model based on ensemble learning strategies was reported by Pereira et al. (2019). Decision trees, logistic regression, and random forests are just a few machine-learning algorithms combined in the model to make predictions.
- Machine learning approaches were utilized by Kemper et al. (2020) to predict student dropout in higher education. A random forest model had the best performance,

according to comparisons made between it and various other models, including logistic regression, decision trees, and neural networks. A machine learning model to predict student dropout from university courses was presented by Solís et al. (2018) Their model combines academic and demographic characteristics as predictors.

- focus on adopting a multi-task learning technique to predict student success in MOOCs. Their model incorporates predictions for three interconnected tasks, including a prediction for student effort, retention, and - looked into the feature that affects students' choices to abandon online. They discovered many variables, including course design, motivation, and academic preparation.

- The dropout predictors in MOOCs were the subject of a systematic literature review undertaken by Alario-Hoyos et al. (2017). They discovered many determinants, such as demographic variables, student participation, and course structure. To predict MOOC dropouts, Chen et al. (2019) compared machine learning techniques, such as logistic regression, decision trees, and random forests.
- Pardos et al. (2013) investigated how affective states throughout the school year predict end-of-year learning outcomes. Their study found that affective states such as boredom, frustration, and confusion strongly predict learning outcomes. Alhramelah and Alshahrani (2020) proposed a machine learning model to predict student dropout in a blended learning environment. Their model uses features extracted from online behavior data, such as clickstream and discussion forum data.
- Finally, Wang and Wang (2019) reviewed the state-of-the-art predicting dropout behavior in massive open online courses. They identified many challenges and opportunities in this field, including the need for more comprehensive data collection for using contextual data to improve predictions.

Our contribution lies in our ability to understand the causes of dropout, as demonstrated by the comparison in Table 1. This is achieved through rich data, which encompasses both sociodemographic and behavioral aspects, ensuring both quality and quantity. Machine learning methods enable us to anticipate which students are at risk of dropping out. However, more than simply detecting students who may drop out is required. Therefore, it is crucial to have an interventionist who can implement a plan to monitor and support at-risk students in their efforts to obtain a diploma.

3 Dataset

A large dataset from a Canadian distance learning platform that was in use from 2016 to 2023 was used in the study. This platform presents a variety of subjects such as mathematics, French, and English, offering students the opportunity to study courses that are tailored to their needs and academic level. It specifically addresses the needs of distant students with a focus on interactive and adaptive learning methods. The dataset in question, which focuses on students enrolled, consists of 35,000 samples and 49 features. This dataset is a useful resource because of the variety of courses offered and the creative online learning model. The output variable has two values: 1

Table 1 Description of student attributes

Aspect	Our Contribution	Cited Research
Data Source	Our used platform with vast and heterogeneous datasets containing socio-demographic and behavioral data	Various learning platforms, student profiles, and prediction methods for identifying dropouts in different online learning environments (Prenkaj et al., 2020).
Feature Selection	Demographic information, past academic performance, and engagement with learning materials	Different combinations of demographic, academic (Bonifro et al., 2020), and behavioral features (Wang & Wang, 2019) to predict dropouts (Alario-Hoyos et al., 2017).
Prediction Method	XGBoost, logistic regression, and random forest methods; accuracy rate of approximately 83% on unseen data	Machine learning techniques such as deep learning, support vector machines, decision trees, random forests, logistic regression, and ensemble learning strategies (Chen et al., 2019; Pardos et al., 2013; Alhramelah and Alshahrani, 2020)
Model Evaluation	ROC curves, confusion matrices, and commonly used performance metrics	Comparisons between different models or investigations of specific aspects, such as affective states, course design, or motivation (Prenkaj et al., 2020)
Practical Implications	Identifying at-risk students, followed by intervention from an interventor to support students in obtaining a diploma and succeeding in an ever-changing job market. Taking into account student motivation problems and providing support to help them persevere in their studies	Identifying at risk students and implementing relevant intervention strategies in different learning contexts or by focusing on specific factors that contribute to dropout (Pereira et al., 2019)

for active students or 0 for students who dropped off the platform. The data comes from various sources, such as student enrollment records, deduced information, historical files, and activity responses. The data can be organized into the following categories: socio-demographic data and behavioral data about the students. The study focuses on students from 2021 to 2023.

- **Socio-demographic data:** This information is fundamental and largely determines each person's social situation, which can influence student behavior and attitudes towards our data includes information about age, gender, ethnic origin, education level, marital status, academic level, geographic location, and parental information

(education level, income, etc.). This data allows us to understand each student's social situation better.

- **Behavioral data:** In the context of online learning on the platform, our behavioural data includes information about the student's interactions with the platform, such as login frequency, duration of each session, courses completed, and the device used (phone, computer, or tablet). This data provides us with a complete picture of how the student uses the platform and will be used to improve student behavior and enhance our predictions' accuracy.

As presented in the following sections, the data will identify patterns and features related to student perseverance, school dropout, and other attributes that may affect the student's academic success.

4 Data exploration

Exploratory data analysis is an essential step before modeling and prediction. This initial exploration aims to provide an initial picture of the points of interest in our dataset.

These analyses help us understand patterns and relationships in data, which are important for predictive analysis. Data exploration requires the use of statistical techniques to understand the characteristics and relationships of data, such as univariate analysis (studying each variable separately), bi-varied research (studying the effects of two variables on a third), and multivariate analytics (studying the simultaneous development of different variables).

We used a correlation matrix to measure the linear relationship between different variables. This helps determine which variables are most closely related and, therefore, the most important for future modeling. We check for strong correlations, weak or non-existent correlations, aberrant values, and patterns.

Values in the correlation matrix range from -1 to 1, a value close to 1 indicating a strong positive correlation between the variables. The feature selection process was conducted using a correlation heatmap. The correlation heatmap can be useful for identifying important features for a predictive model and optimizing the feature selection process. The strongly correlated features were used for model building due to their high impact on student outcomes. The dataset was normalized using a standard scaler to eliminate the mean and scale it to unit variance in the final data preprocessing stage.

The correlation matrix shown in Fig. 1 of the dataset analysis provided significant insights into the underlying patterns. Notably, a robust positive correlation was observed between `weekly_minutes` and `active_days`, as well as with the time spent on the platform over the past month, three months, and six months (`heures_derniers_mois`, `heures_derniers_3_mois`, `heures_derniers_6_mois`), suggesting that consistent engagement over time is indicative of sustained student activity. Conversely, demographic factors such as `born_outside_canada`, `did_study_in_quebec`, and educational background `has_des` showed negligible correlations with other variables, highlighting

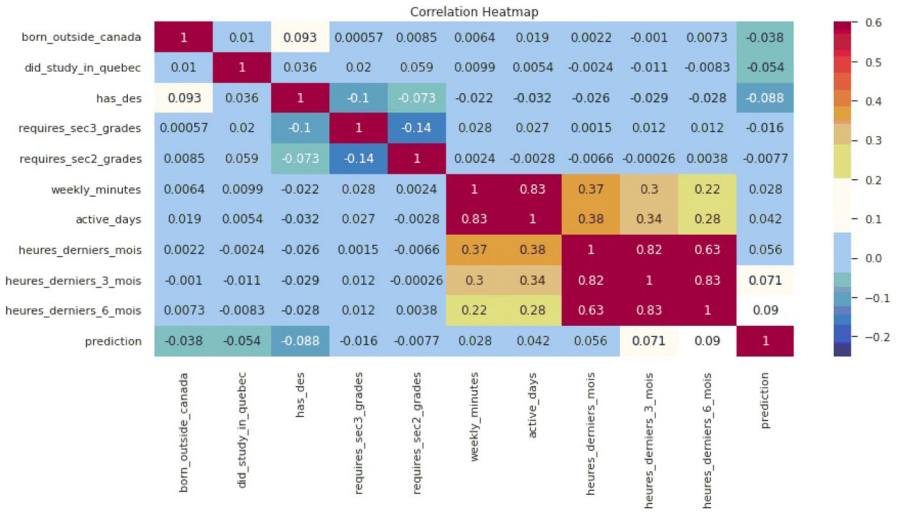

Fig. 1 Correlation matrix for revealing interrelationships among attributes

their limited linear influence on the measured engagement metrics. The outcome variable `prediction` exhibited no strong linear correlations with the features, indicating that single feature may not be significant predictors of student retention, thereby suggesting the need for more complex modeling to understand the determinants of student engagement and success within this online educational context. The correlation analysis of data revealed a strong positive correlation between weekly minutes spent and active days, as well as the time invested on the platform over varying durations, indicating persistent student activity. Demographic and educational factors had minimal impact on engagement metrics. There was a lack of correlation observed with the outcome variable ”prediction,” indicating that complex models are needed for predicting keeping students engaged.

5 Methodology

Artificial intelligence (AI) has become increasingly influential in various areas of our lives, providing new tools and services that help decision-making. This technological advancement has significantly impacted the education sector as we can augment multiple aspects of the learning process by applying AI (Artificial Intelligence) and ML (Machine Learning) algorithms (Issah et al., 2023). One of the key advantages of these technologies is their capacity to create predictive models, which can help students effectively plan their future and boost their academic performance. This potential has led to a surge in interest in employing AI and ML in education to support student perseverance.

Our approach in Fig. 2 uses widely recognized machine learning algorithms, which offer an efficient system for school dropout prediction. We use various machine learning methods on our data to predict which students will be active or not. The objective

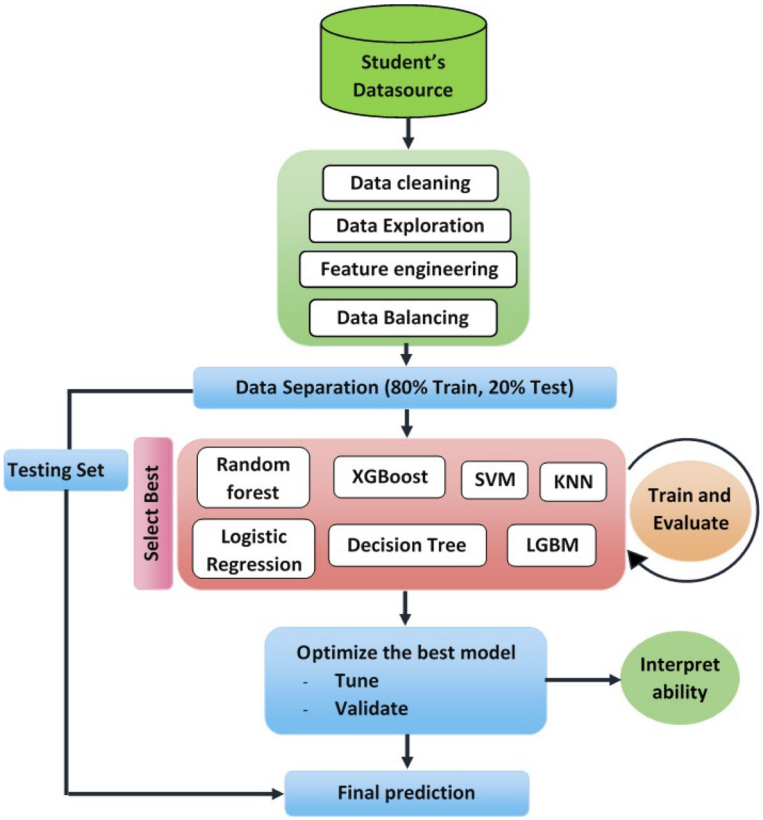

Fig. 2 Process of school dropout prediction

is to create a model based on the students' current accomplishments and activities to anticipate their activity level. A binary classifier can address a common classification problem to predict whether a student can complete the activities.

Classification techniques in machine learning are explored for identifying appropriate predictive models. The main idea is to find the most suitable prediction model and compare each applied classifier's various prediction performance metrics. Hyperparameters were tuned using the grid search technique, which is widely used for finding the best settings of parameters.

Student profiling

In the context of our data, clustering on Fig. 3 is used to group students based on similar characteristics or behaviors that may expose them to a risk of dropping out. By categorizing students based on their interactions and conduct on the platform, we hope to discover hidden student profiles.

Student profiling has allowed us to discover our dataset's two main student profiles (dropouts and active students). In addition, a student can have a dropout or active profile depending on their status, which can take on different values such as dropout-inactivity, dropout-third-enrollment, etc.

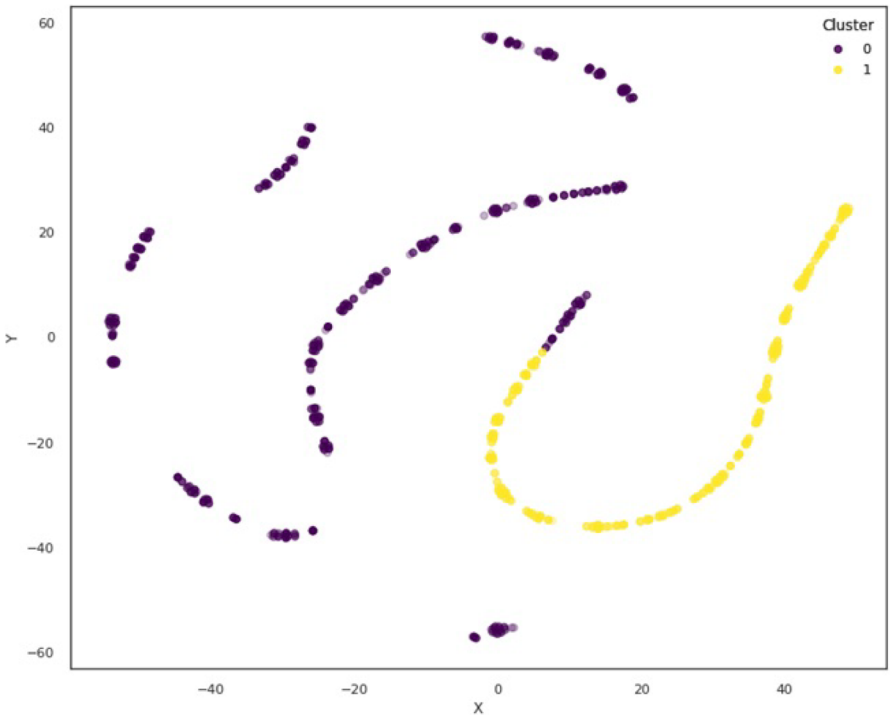

Fig. 3 Students profils clustering

Prediction

Prediction refers to using machine learning algorithms to predict the future behaviors of students based on observed sociodemographic and behavioral data; the process is determined in the algorithm 1 as it is important to rank the attributes for suitable feature selection.

Attribute importance evaluation

Due to the many attributes that describe our platform and students, it is important to conduct a study to determine the most relevant characteristics for effectively using this data. It is possible to decide on the ideal.

• RFECV (Recursive Feature Elimination with Cross-Validation) investigates the suitable number of features for a machine learning model, intending to attractively eliminate elements and select the feature set that optimizes model performance based on cross-validation. The RFECV plot in Fig. 4 for the LGBMClassifier demonstrates the optimization of feature selection to maximize model accuracy. The optimal feature count was determined to be 31, which yielded a peak score of 0.831, indicating that beyond this number, additional features do not contribute to a significant increase in model performance. This illustrates the effectiveness of RFECV in identifying a subset of features that are most predictive while avoiding model overfitting due to excessive complexity.

Algorithm 1 Student Dropout Prediction.

Require: Dataset D
$D' \Leftarrow C(D)$ ▷ $C(D)$ represents the data cleaning function
$features\_ranking \Leftarrow SHAP(D')$ ▷ $SHAP$ measures the importance of each feature
$D'' \Leftarrow Select\_Features(D', features\_ranking)$
$k \Leftarrow 2$ ▷ Define the number of clusters as 2: 'DropOut' and 'Active'
$clusters \Leftarrow Kmeans(D'', k)$ ▷ Perform k-means clustering on D''
for each student i in D'' **do**
  if $clusters[i] == 0$ **then**
    Label student i as 'DropOut'
  else
    Label student i as 'Active'
  end if
end for
$XGModel \Leftarrow Train\_XGBoost(D'')$
$RFModel \Leftarrow Train\_RandomForest(D'')$
$LRModel \Leftarrow Train\_LogisticRegression(D'')$
for each student i in D'' **do**
  $P\_XG[i] \Leftarrow XGModel(D''[i])$
  $P\_RF[i] \Leftarrow RFModel(D''[i])$
  $P\_LR[i] \Leftarrow LRModel(D''[i])$
  $P[i] \Leftarrow \frac{P\_XG[i]+P\_RF[i]+P\_LR[i]}{3}$ ▷ Simple average ensemble of the three models
  if $P[i] > 0.5$ **then**
    Predict student i as 'Likely to DropOut'
  else
    Predict student i as 'Likely to be Active'
  end if
end for

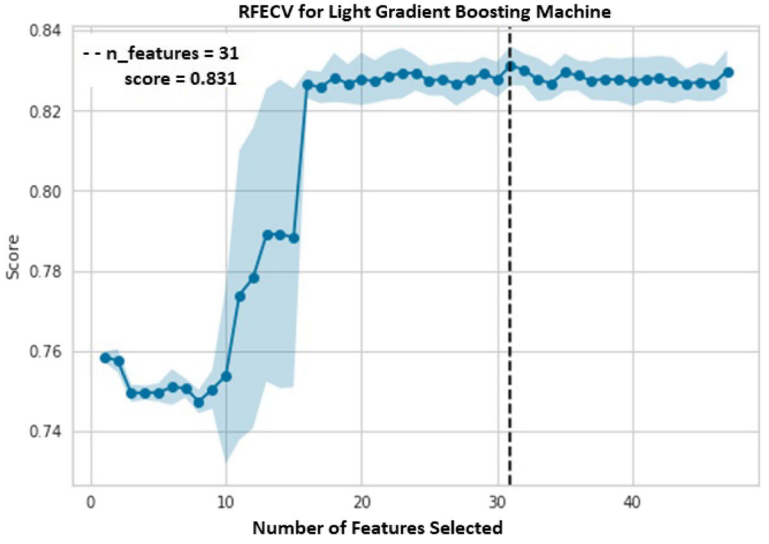

Fig. 4 Selecting the optimal number of attributes

Although the number of attributes was determined, another study was carried out to rank the importance of each point. To calculate the contribution of each feature to the final prediction, SHAP (Shapley Additive Explanations) was chosen as the method for determining the significance of each attribute in a machine-learning model for the platform.

Each SHAP value in Fig. 5 represents the expected impact of a feature on the model output, assuming all other attributes have their mean values. In the image mentioned, the variables with a high positive contribution when their values are high, and a low negative contribution when their values are common, can be observed. All features are explained in Table 2 and presented in order of importance for the overall feature selection. The first feature is the most important, while the last is the least important."

Our study is based on the XGBoost algorithm and compared with the following algorithms: logistic regression and random forest. To apply these machine learning techniques to our data, we first define the target feature (e.g., dropout) and split the data into training and test sets. We then train the model on the training data, using the most frequent class as the prediction. Additionally, we applied the SMOTE (Synthetic Minority Oversampling Technique) method for synthetic oversampling of the minority class, as dropouts represent a minority in our dataset.

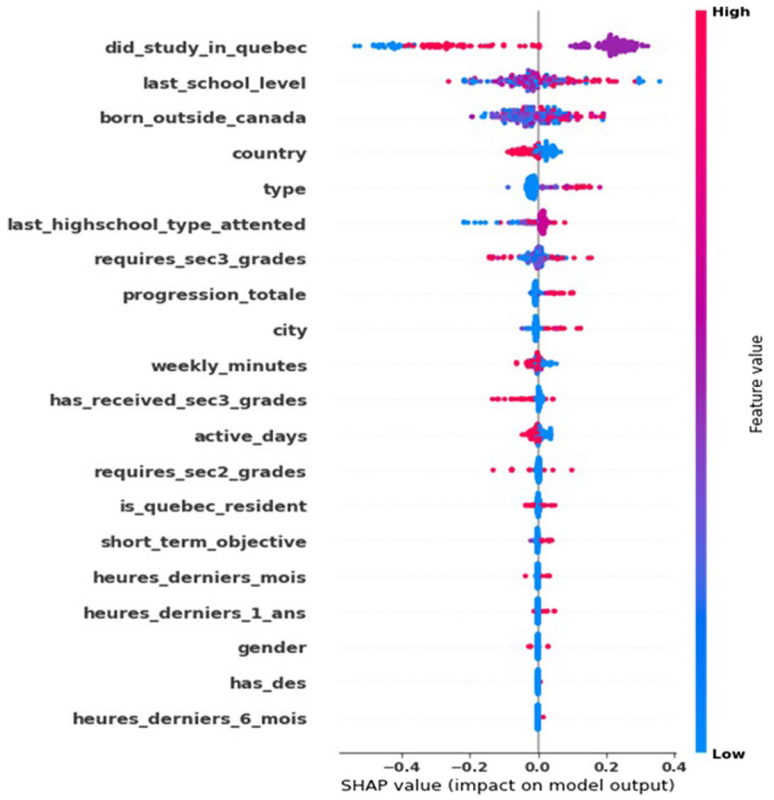

Fig. 5 Ranking the attribute importance

Table 2 Description of student attributes

Attribute	Description
Did study in Canada	Whether the student has studied in Canada
Last school level	The highest level of education the student has completed
Born outside Canada	Whether the student was born outside of Canada
Country	The country in which the student currently resides
Type	The type of educational program or institution the student is attending
Last high school	The name of the last high school the student attended
Require sec3 grade	Whether the student requires a specific grade in secondary 3 for their current educational goals
Total progress	The total progression of the student in their current educational program
City	The city in which the student currently resides
Weekly minutes	The number of minutes the student spends on educational activities per week
Has received sec3 grade	Whether the student has received a specific grade in secondary 3
Active days	The number of days the student has been active in their current educational program
Require sec2	Whether the student requires a specific grade in secondary 2 for their current educational goals
Is Quebec resident	Whether the student is a resident of Quebec
Short term objective	The student's short-term educational or career objectives
Hours last months	The number of hours the student spent on educational activities in the last month
Hours last years	The number of hours the student spent on educational activities in the last year
Gender	The gender of the student
Has DES	Whether the student has completed a Diplôme d'études secondaires (DES)
Hours last 6 months	The number of hours the student spent on educational activities in the last 6 months

We addressed the issue of overfitting during model training by adopting the SMOTE method. SMOTE helps to overcome imbalanced data, where the number of instances in the dropout class is significantly lower than the other class (active students), by generating synthetic instances of the minority class.

To address the issue of an imbalanced dataset, we used a sampling method to improve the accuracy of our prediction process, which is preferable to applying SMOTE. We then evaluate the models' performance on the test data using metrics such as accuracy, precision, recall, and F1 score.

XGBoost (Extreme Gradient Boosting) (Chen & Guestrin, 2016) is a machine learning algorithm that uses gradient boosting to train decision trees. It is a type of adaptive learning where multiple models are combined to improve the accuracy of

predictions. XGBoost is a suitable model for various applications like classification, regression, and ranking problems.

Logistic regression (King & Zeng, 2001) is a model used to predict our binary target variable (dropout or non-dropout) from several independent attributes. Using our data and profiling users, it is possible to employ logistic regression to predict the dropout risk among students based on features such as their sociodemographic profile, behavior on the platform, and relationships with teachers.

Random Forest (Pal, 2005) is a powerful and robust machine learning model we integrated to predict dropout. It is called a "forest" because it involves creating multiple trees (hence the name forest) and using them to make predictions. The main idea of a random forest is to combine the predictions of various decision trees to boost the collective performance of the model. One way to achieve this is to average the predictions of each tree, but other methods can also be used.

The choice between the models used was based on the specifics of our data, as our dataset has many attributes and prediction goals.

6 Experimental results and discussion

In this work, our goal is to accurately predict student dropout on our platform using a Python-developed approach with the sci-kit-learn package. We utilize a combination of XGBoost, Random Forest, and logistic regression (LR) for this purpose. The input data was split into training (80%) and testing (20%) datasets, a proportion chosen to balance substantial training data with enough testing data to validate the model's predictive power and minimize overfitting risk. Increasing the training data beyond 80% may improve performance on that data but risk overfitting and reduce generalizability, while increasing the testing data beyond 20% can enhance confidence in generalizability but might compromise the model's effectiveness due to a smaller and potentially less representative training set.

Our rigorous multi-step methodology involves data preparation (see Section 4), model training, hyperparameter tuning, and validation. The robust training dataset allows comprehensive pattern evaluation, while the testing dataset assesses the models generalizability. Figure 6 illustrates the learning curve for an XGBClassifier model, where the training score decreases and the cross-validation score increases with more training instances, indicating improved model generalization.

Hyperparameter tuning is crucial for optimizing model performance. We employ techniques like grid search and random search to fine-tune hyperparameters, including learning rate, max depth, `n_estimators`, subsample, gamma, and `reg_alpha`. This tuning enhances the XGBoost model's ability to detect school dropouts effectively in binary classification.

Our analysis, as shown in Table 1, compares our model with similar predictive models used in practice. This comparison not only highlights our method's novel features but also establishes a baseline for performance evaluation.

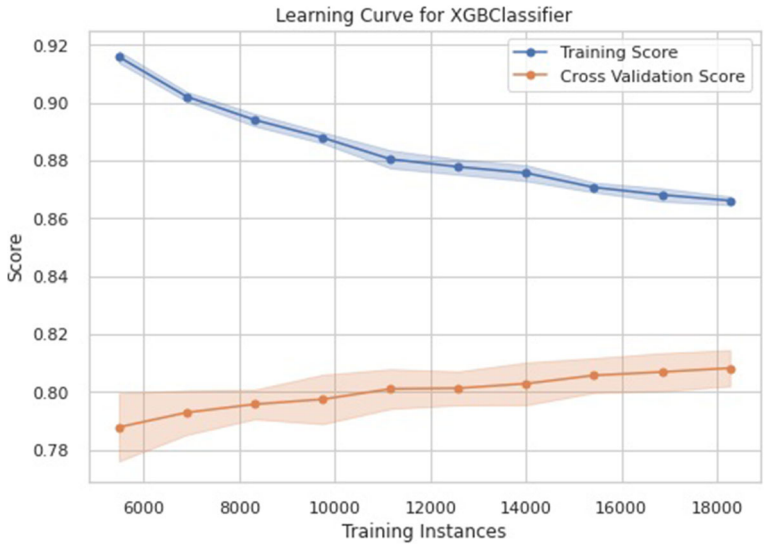

Fig. 6 The learning curve of XGboost

6.1 Comparison of the prediction performance

Based on the Fig. 7, which show the variation of accuracy, precision, recall, and F1 score, the top-performing models are the Decision Tree Classifier and the Light Gradient Boosting Machine, with an accuracy score of 0.7981. However, it is essential to consider other metrics as well, such as the area under the curve (AUC), recall, precision, F1 score, Cohen's kappa, Matthews correlation coefficient (MCC), and training time (TT).

By examining the AUC metric, which measures how well the model can distinguish between positive and negative cases, the Light Gradient Boosting Machine is the top performer with a score of 0.8625. It is followed closely. By the Random Forest Classifier and the Extra Trees Classifier, with scores of 0.8620 and 0.8426, respectively.

Regarding recall, which represents the proportion of actual positive cases the model correctly identified, the Extra Trees Classifier is the top performer with a score of 0.7783. It is followed by the Light Gradient Boosting Machine and the Decision Tree Classifier, with scores of 0.7743 and 0.7583, respectively.

For precision, which indicates the proportion of positive cases correctly identified by the model, the Decision Tree Classifier is the top performer with a score of 0.7482. The Light Gradient Boosting Machine and the Extra Trees Classifier follow closely behind, with scores of 0.7456 and 0.7258, respectively.

Regarding the F1 score, which is the harmonic mean of recall and precision, the Light Gradient Boosting Machine is the top performer with a score of 0.7548. The Decision Tree Classifier and the Extra Trees Classifier come next with scores of 0.7508 and 0.7497, respectively. For Kappa, which measures the agreement between the predicted and actual labels, the Light Gradient Boosting Machine is the top performer with a score of 0.5840, followed by the Decision Tree Classifier with a score of 0.5814.

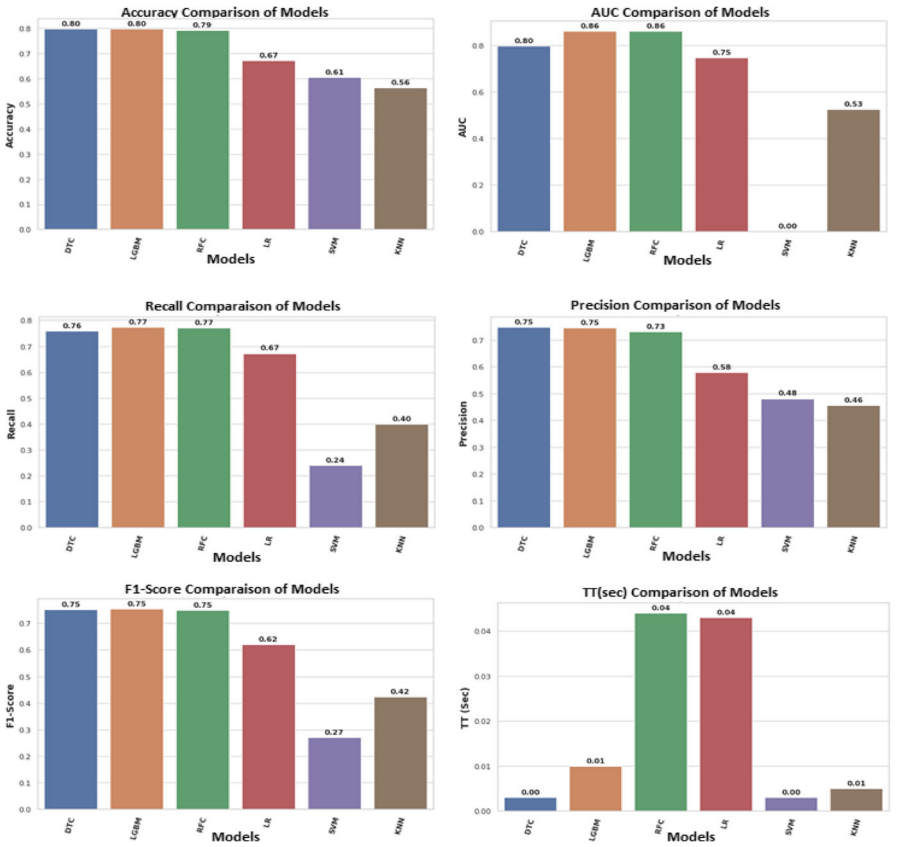

Fig. 7 Metrics evaluation performance of the prediction model

When considering the MCC (Matthews Correlation Coefficient), which measures the correlation between predicted and actual labels, the Light Gradient Boosting Machine comes out on top with a score of 0.5894. It's closely followed by the Decision Tree Classifier, which achieves a score of 0.5839. Finally, the fastest model for training time is the Decision Tree Classifier, with a time of 0.0030 seconds. The linear discriminant analysis, the ridge classifier, and the SVM with a linear kernel follow closely behind, each taking 0.0030 seconds for training. The CatBoost classifier has the longest training time at 1.8910 seconds.

In summary, the XGBoost model appears to be the best overall performer, with the highest AUC score, second-highest accuracy score, third-highest recall score, second-highest precision score, highest F1 score, highest Kappa score, and highest MCC score, as well as a reasonable training time. However, the Decision Tree Classifier may also be a good option, with the highest precision score, the second-highest recall score, the third-highest F1 score, and the fastest training time. Ultimately, the model chosen will depend on the specific requirements and limitations for predicting school dropout.

In our binary classification problem, we used our dataset to determine if a student is likely to drop out or stay active on the platform. We created ROC curves and evaluated their properties to determine the effectiveness of our model. From a selection of prediction models, we identified the three major predictors of school dropout.

Specifically, the ROC curve represents a two-dimensional graph in which the true positive rate (TPR) is denoted on the y-axis. In contrast, the false positive rate (FPR) is represented on the x-axis.

This example shows that the decision tree performs well, as it has a high TPR and a low FPR across a range of different classification thresholds. The area under the curve (AUC) is also quite high, indicating that the classifier can distinguish between the two classes with high accuracy.

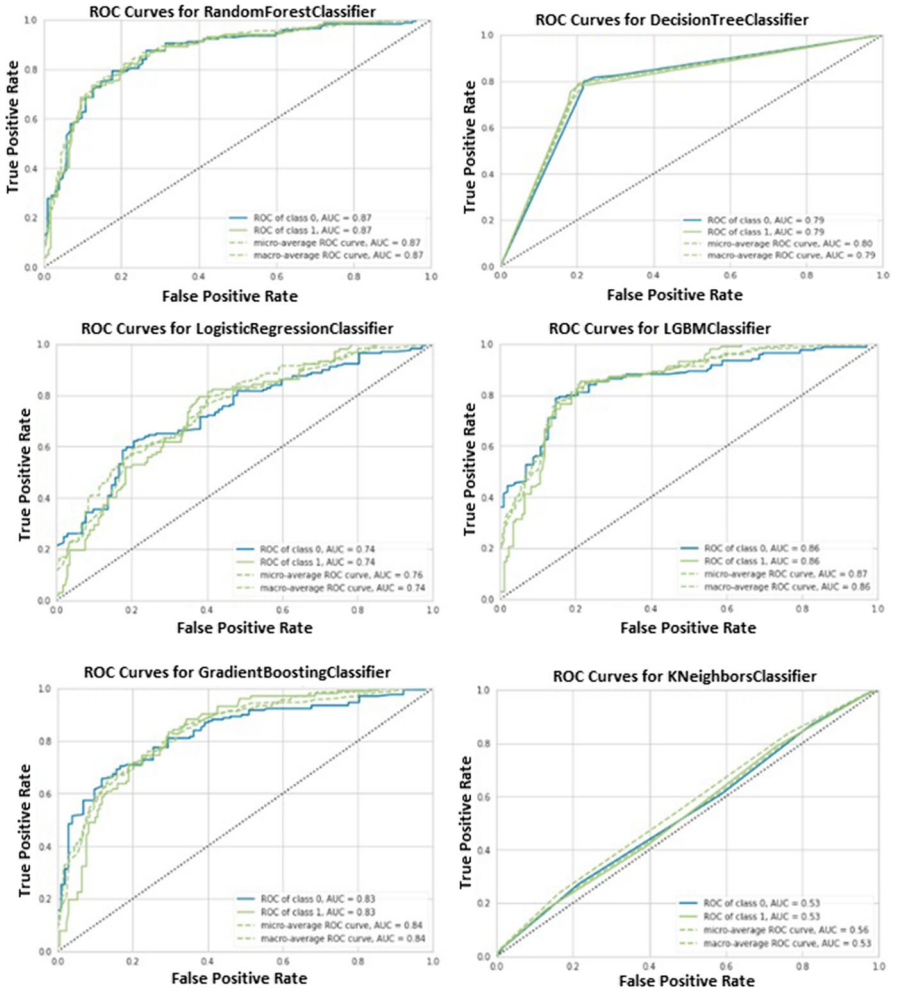

Fig. 8 Amplified ROC curves visualization for detailed analysis

Overall, interpreting a ROC curve for a decision tree involves looking at the TPR and FPR values at different thresholds and assessing the classifier's overall performance. The blue curve represents the ROC curve of the classifier, and the dotted line represents the performance of an XGBoost model.

As we can see, the XGBoost model can perform significantly better than a random classifier and can help assess how well the classifier distinguishes between the two classes (dropout or active).

The Fig. 8 shows that the AUC results and the corresponding ROC curves for the Random Forest (RF) model are 0.82. A higher AUC score obtained from our model implies better and acceptable classification performance because points representing

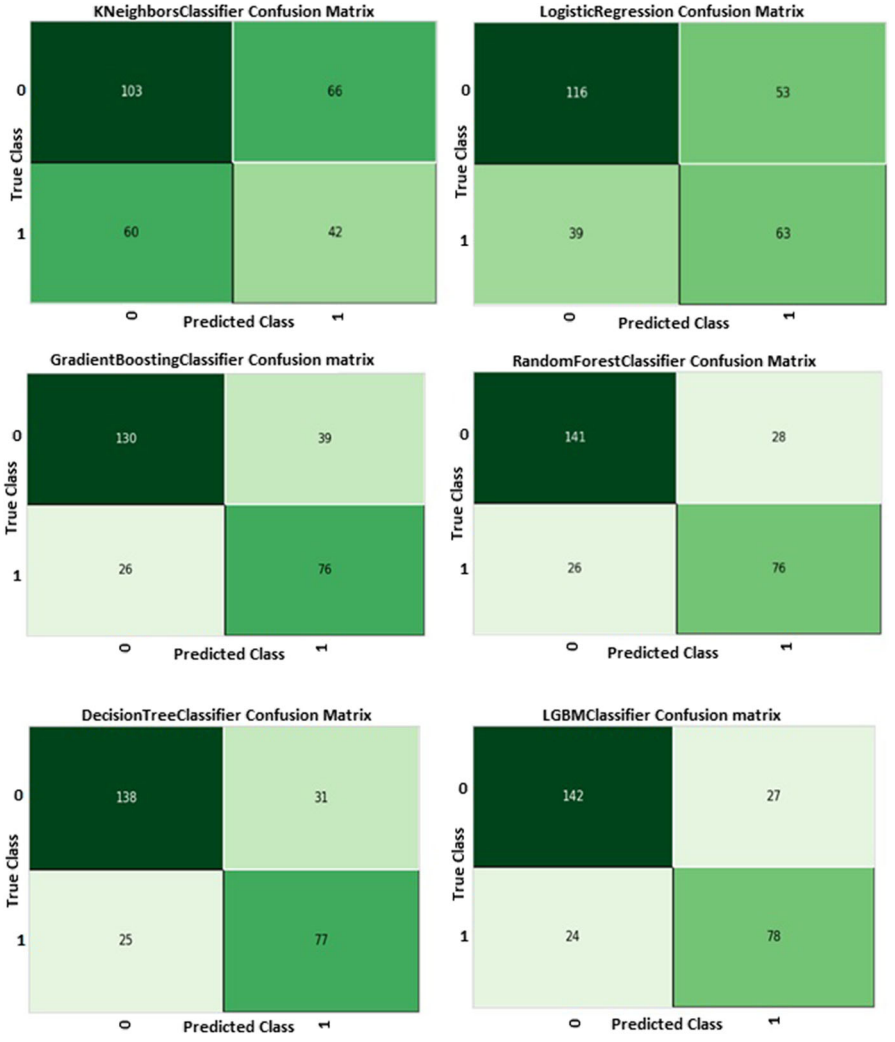

Fig. 9 Evaluation of model predictions via confusion matrix

model classification better than random guesses are located above the diagonal line. According to the discussion of the performance metrics, the selected binary classification models may accurately predict whether or not students will leave the platform. However, various performance metrics must be investigated to ensure the classifiers' performance.

In analyzing the performance of the six classifiers through their confusion matrices, we observe varied predictive abilities. For instance, the Gradient Boosting classifier and Random Forest classifier exhibits a strong predictive accuracy for true negatives, suggesting a high capability to correctly identify students who will continue their studies. On the other hand, the Logistic Regression classifier presents a more balanced approach, with relatively equal numbers of true positives and true negatives, indicating a consistent performance in both identifying dropouts and non-dropouts.

Each matrix in Fig. 9 has two classes (0 and 1) and provides four pieces of information: the number of true negatives (top left), false positives (top right), false negatives (bottom left), and true positives (bottom right). These matrices are key in evaluating the classification performance of the different models.

These measures allow us to analyze the XGBoost classifier's capabilities for predicting dropout cases and its drawbacks. For instance, with an accuracy of 82 %, the model can predict the correct outcome in nearly 80 % of the cases. However, the precision of 71.3 % indicates that about 30 % of the predicted dropouts are false positives, which could result in incorrect interventions or student recommendations. Similarly, the recall of 75.5% shows that the model can recognize about 75% of the actual dropouts but misses 25%, which may lead to missed opportunities for intervention. This quantitative analysis informs the refinement of these models for better dropout prediction Krüger et al. (2023).

7 Conclusion

Distance learning platforms are innovative and accessible solutions that provide diverse learning options to close the skills gap and respond to students' evolving demands. Predictive models play a crucial role in understanding student engagement and outcomes. To this end, we have investigated various features and modeling strategies, including data preprocessing, feature extraction methods, modeling algorithms, and model evaluation methods. Our approach explores the potential of different prediction models to improve the identification of prospective dropouts. While our analysis of a widely used prediction model achieved an accuracy rate of 82%, the precision of 73.1% and recall of 75.5% indicate that it can still be improved in recognizing potential dropouts. Improving models of prediction for identifying likely dropouts in distance learning systems is essential for making interventions and recommendations as precise and efficient as possible.

Overall, the combination of innovative platforms and advanced predictive models has the potential to revolutionize distance learning and help students succeed in an ever-changing job market. Understanding the causes of dropout and developing more effective prevention strategies can significantly contribute to students' academic success using distance learning systems.

Funding This work was supported by Ministry of the Economy in Canada.

Data Availability Statement The datasets generated and/or analyzed during this study are not publicly available. This is due to a confidentiality agreement and the fact that the data is hosted exclusively on ChallengeU's server. The data is proprietary to ChallengeU, thus it is not accessible to anyone outside of the company. This study was conducted within the company's servers using this non-public data.

Declarations

Conflicts of interest The authors declare that they have no conflict of interest.

References

Alam, R., Ahmad, N., Shahab, S., & Anjum, M. (2023) Prediction of dropout students in massive open online courses using ensemble learning: A pilot study in postcovid academic session. In: *Mobile computing and sustainable informatics* (pp. 549–565)

Alario-Hoyos, C., Estévez-Ayres, I., Pérez-Sanagustín, M., Kloos, C. D., & Fernández-Panadero, C. (2017). Understanding learners' motivation and learning strategies in moocs. *The International Review of Research in Open and Distributed Learning, 18*, 119–137.

Alhramelah, A., & Alshahrani, H. A. (2020). Saudi graduate student acceptance of blended learning courses based upon the unified theory of acceptance and use of technology. *Australian Educational Computing, 35*, 1–22.

Bonifro, F. D., Gabbrielli, M., Lisanti, G., & Zingaro, S. P. (2020). Student dropout prediction. *Artificial Intelligence in Education, 12163*, 129–140.

Chen, J., Feng, J., Sun, X., Wu, N., Yang, Z., & Chen, S.-S. (2019). Mooc dropout prediction using a hybrid algorithm based on decision tree and extreme learning machine. *Journal Hindawi Mathematical Problems in Engineering*

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*

Issah, I., Appiah, O., Appiahene, P., & Inusah, F. (2023). A systematic review of the literature on machine learning application of determining the attributes influencing academic performance. *Decision Analytics Journal, 7*, 100204.

Kemper, L., Vorhoff, G., & Wigger, B. U. (2020). Predicting student dropout: A machine learning approach. *European Journal of Higher Education, 10*, 28–47.

King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis, 9*, 137–163.

Krüger, J. G. C., Souza Britto, A., & Barddal, J. P. (2023). An explainable machine learning approach for student dropout prediction. *Expert Systems with Applications, 233*, 120933.

Oz, H. C., Güven, Ç., & Nápoles, G. (2022). School dropout prediction and feature importance exploration in malawi using household panel data: machine learning approach. *Journal of Computational Social Science, 6*, 245–287.

Pal, M. (2005). Random forest classifier for remote sensing classification. *International Journal of Remote Sensing, 26*, 217–222.

Pardos, Z.A., Baker, R., Pedro, M. O. S., Gowda, S. M., & Gowda, S. M. (2013). Affective states and state tests: investigating how affect throughout the school year predicts end of year learning outcomes. In: *International Conference on Learning Analytics and Knowledge*

Pereira, F. D., Oliveira, E. H. T., Cristea, A. I., Fernandes, D., Silva, L., Aguiar, G., Alamri, A., & Alshehri, M. (2019). Early dropout prediction for programming courses supported by online judges. In: *International Conference on Artificial Intelligence in Education*

Prenkaj, B., Velardi, P., Stilo, G., Distante, D., & Faralli, S. (2020). A survey of machine learning approaches for student dropout prediction in online courses. *ACM Computing Surveys (CSUR), 53*, 1–34.

scolaire. (2023). https://www.ledevoir.com/opinion/idees/753858/milieux-defavorisesplus-de-10-000-decrocheurs-scolaires-au-quebec

Shiao, Y. -T., Chen, C. -H., Wu, K. -F., Chen, B. -L., Chou, Y. -H., & Wu, T. -N. (2023). Reducing dropout rate through a deep learning model for sustainable education: long-term tracking of learning outcomes of an undergraduate cohort from 2018 to 2021. *Smart Learning Environments, 10*

Solís, M., Moreira, T. M. B., Gonzalez, R., Fernandez, T., & Hernandez, M. (2018). Perspectives to predict dropout in university students with machine learning. *IEEE International Work Conference on Bioinspired Intelligence (IWOBI), 2018*, 1–6.

Wang, L., & Wang, H. (2019). Learning behavior analysis and dropout rate prediction based on moocs data. *2019 10th International Conference on Information Technology in Medicine and Education (ITME)*, 419–423