Computers and Education: Artificial Intelligence 6 (2024) 100205

Contents lists available at ScienceDirect

Computers and Education: Artificial Intelligence

journal homepage: www.sciencedirect.com/journal/computers-and-education-artificial-intelligence

Machine learning model (RG-DMML) and ensemble algorithm for prediction of students' retention and graduation in education

Kingsley Okoye [a], Julius T. Nganji [b], Jose Escamilla [c], Samira Hosseini [a,d,*]

[a] Writing Lab, Institute for Future of Education, Tecnologico de Monterrey, Monterrey, CP 64849, Nuevo Leon, Mexico
[b] Department of Occupational Science & Occupational Therapy, University of Toronto, Canada
[c] Institute for Future of Education, Tecnologico de Monterrey, Monterrey, CP 64849, Nuevo Leon, Mexico
[d] School of Engineering and Sciences, Tecnologico de Monterrey, Monterrey, CP 64849, Nuevo Leon, Mexico

ARTICLE INFO

Keywords:
AI in education
Machine learning
Predictive modeling
Educational data
Educational innovation
Classification algorithm
Supervised learning

ABSTRACT

Automated prediction of students' retention and graduation in education using advanced analytical methods such as artificial intelligence (AI), has recently attracted the attention of educators, both in theory and in practice. Whereas invaluable insights and theories for measuring and testing the topic have been proposed, most of the existing methods do not technically highlight the non-trivial factors behind the renowned challenges and attrition. To this effect, by making use of two categories of data collected in a higher education setting about students (i) retention ($n = 52262$) and (ii) graduation ($n = 53639$); this study proposes a machine learning model - RG-DMML (retention and graduation data mining and machine learning) and ensemble algorithm for prediction of students' retention and graduation status in education. This was done by training and testing key features that are technically deemed suitable for measuring the constructs (retention and graduation), such as (i) the Average grade of the previous high school, and (ii) the Entry/admission score. The proposed model (RG-DMML) is designed based on the cross industry standard process for data mining (CRISP-DM) methodology, implemented using supervised machine learning technique such as K-Nearest Neighbor (KNN), and validated using the *k*-fold cross-validation method. The results show that the executed model and algorithm based on the Bagging method and 10-fold cross-validation are efficient and effective for predicting the student's retention and graduation status, with Precision (retention = 0.909, graduation = 0.822), Recall (retention = 1.000, graduation = 0.957), Accuracy (retention = 0.909, graduation = 0.817), F1-Score (retention = 0.952, graduation = 0.885) showing significant high accuracy levels or performance rate, and low Error-rate (retention = 0.090, graduation = 0.182), respectively. In addition, by considering the individual features selected through the Wrapper method in predicting the outputs, the proposed model proved more effective for predicting the students' retention status in comparison to the graduation data. The implications of the models' output and factors that impact the effective prediction or identification of at-risk students, e.g., for timely intervention, counselling, decision-making, and sustainable educational practice are empirically discussed in the study.

1. Introduction

Today, predicting students' retention and graduation status has become one of the most challenging and important data management and investment priorities for educators (Arqawi et al., 2022; Dake & Buabeng-Andoh, 2022; Delen, 2010; OECD, 2019, 2022a; UNESCO, 2020, 2022a). Globally, the Organization for Economic Corporation and Development's (OECD) data on individual students who enter a degree program in higher education (OECD, 2019; 2022a) shows that 39 percent (%) of learners graduate within the estimated time of the program. While 12% of full-time equivalent of the learners are likely to drop out before their second year, 24% after three years, and 20% by the end of the individual program. Demographic -wise, among the different economies and countries, there are also disparities in the students' retention and graduation rate. Ranging from less than 30 percent (%) in most economies to 60% in some countries in Europe (OECD, 2019; 2022a). Globally, all countries and economies according to the OECD's data on students' retention and graduation, increase by more after three years of the students participating in the various individual programs, especially in programs where the theoretical duration is lesser (OECD,

* Corresponding author. Writing Lab, Institute for Future of Education, Tecnologico de Monterrey, Monterrey, CP 64849, Nuevo Leon, Mexico.
E-mail address: samira.hosseini@tec.mx (S. Hosseini).

https://doi.org/10.1016/j.caeai.2024.100205
Received 4 September 2023; Received in revised form 8 January 2024; Accepted 11 January 2024
Available online 12 January 2024

2019; 2022a). For example, countries like New Zealand, Netherlands, and Switzerland show an increase of 40%, Brazil at 50%, and the UK at 85%, respectively (OECD, 2019; 2022a). Another factor/feature that has been shown to impact the students' retention or graduation rate is the upper secondary (high school) degree or vocational upper programme status (Brdesee et al., 2022; Dake & Buabeng-Andoh, 2022; Delen, 2010; OECD, 2019, 2022a). There is evidence that 38% of upper secondary, and 35% of vocational upper secondary students who enter a tertiary or higher education program, graduate within the estimated time of the program (OECD, 2019; 2022a). Moreover, those figures are estimated by the OECD (2019, 2022a) to increase by 12% or more in three years of not completing the program within the theoretical duration.

In this study, we note that having the capability to predict the rate or status of the learners' retention and graduation based on the students' (educational) generated data can enable the many institutions or educators to take preventive measures, and also support the educational management, including informative research and development studies of this type (Arqawi et al., 2022; Brdesee et al., 2022; Cardona & Cudney, 2019; Dake & Buabeng-Andoh, 2022; Delen, 2010; Muncie, 2020; Nayak et al., 2023; Okoye et al., 2022; Palacios et al., 2021; Ploutz, 2018; Uliyan et al., 2021). Data mining (DM) or yet machine learning (ML) approaches can be employed to extract valuable information and to detect useful patterns from the student databases, by developing automated and intelligent models that can predict students at risk of dropping out at different levels of their study, or delayed to no graduation (Arqawi et al., 2022; Muncie, 2020; Palacios et al., 2021; Uliyan et al., 2021). The outcome of the methods (AI-based) can help offer a welcome plan or support program for at-risk students who are identified by the predictive models or applications, to assist in improving the academic results or educational processes, establish a data monitoring plan to track the academic status of the students, and for further analysis or educational decision-making (Palacios et al., 2021). Technically, the goal of the data mining or AI/ML studies and applications toward the students' retention and graduation status detection, can be attributed to the analytics or process modeling methods that are heuristically used to predict and explain the different factors, sensitivity, or reasons behind the student's attrition, delayed to no graduation, or aiding of the decision-making processes by the educators (Delen, 2010; Ploutz, 2018; Priyambada et al., 2023). The impact and holistic applications of such type of research and process modeling method are done by making use of suitable feature selection mechanisms, e.g. the Wrapper method applied in this study (Maldonado et al., 2022), to model and detect key features (attributes) that purportedly contribute to the students' attrition, including an understanding of non-trivial factors behind the individual feature selection and techniques, and building of the so-called students' attrition predictive model (AI systems) using the machine learning methods (supervised or unsupervised) with the highest level of classification accuracy or performance, as demonstrated in this study.

1.1. Rationale of the study

Our review of the literature shows that the scientific or documented evidence has become clearer on the need for studying various factors or features that can be used to predict the students' retention or graduation rate in education (OECD, 2022a). Previous research has shown that to improve students' retention and graduation, educators must try to understand the non-trivial factors behind the attrition phenomenon (Delen, 2010; Priyambada et al., 2023). There is a need for data mining methods or AI approaches for identifying students who may be at risk of dropping out of school or not graduating within the expected period (Priyambada et al., 2023). In the past, studies have investigated the social and didactic dimensions of the students' attrition, and have revealed invaluable insights and theories for measuring and testing of the same (Delen, 2010; Malik, 2011; Veenstra, 2009). However, they do not provide the accentuated and much-needed/listed requirement for forecasting and accurate prediction of the non-trivial factors behind the attrition and challenges. Only recently has research in the prediction of students' retention or graduation through technical data requirements or analytical means attracted much attention in the literature (Arqawi et al., 2022; Cardona & Cudney, 2019; Dake & Buabeng-Andoh, 2022; Muncie, 2020; Ploutz, 2018; Priyambada et al., 2023; Uliyan et al., 2021).

Along these lines, by making use of two categories of data collected within the higher education setting about students (i) Retention and (ii) Graduation; this study proposes a machine learning model - RG-DMML (retention and graduation data mining and machine learning) and ensemble algorithm through Bagging technique and Wrapper method for feature selection (LaViale, 2023; Maldonado et al., 2022; Ngo et al., 2022; Zhang et al., 2019) for prediction of the students' retention and graduation status in education. This is done by investigating the key features or parameters that are deemed suitable for measuring the two constructs (retention and graduation) (OECD, 2019; 2022a). The implications of the results and model outputs, and how those are associated to the students' retention and graduation, especially toward sustainable educational practice are also discussed in the paper.

The main research questions of this study are:

1. How can we develop machine learning model and algorithm that prove effective for the prediction of students' retention and graduation rate in education?
2. What features are more suitable for predicting the students' retention and graduation status with a high level of performance and accuracy?

The main contributions of this study include:

1. It introduces a machine learning model (RG-DMML) that proves effective and efficient for the prediction of students' retention and graduation status in education.
2. It describes key features that are suitable and more effective for the prediction of the students' retention and graduation with a high level of performance.
3. It defines an ensemble algorithm applied for the implementation of the RG-DMML model with a high level of classification or accuracy.
4. It empirically discusses both the technical and didactical implications of the results and model output toward a sustainable educational practice, particularly as it concerns the students' retention and graduation in education.

2. Background information

2.1. Students' retention and graduation in education

Research has shown that the impact or effectiveness of educational systems can be determined by examining the students' completion rate or path that they follow once they enter higher education (OECD, 2019; 2022a). Different factors may impact the students' pathways through higher education. The OECD data (2019, 2022a) on student completion rate in tertiary education shows that on average 12% of students who start a degree program end up dropping out, 2% or more having to transfer to another education level, and 85% still enrolled in the same or another program. Interestingly, a large proportion of those transfers between the higher education levels, mainly occur soon after the student has entered or enrolled in a program (OECD, 2019). The gap between the student' drop-out after their first year of study is also considerably broader in some countries than in others. Ranging from 6% in USA, to around 20% in some countries in Europe. Those figures compared to the past, show that the rate of students who leave the educational system, particularly higher education, without graduating has considerably increased over time (Dake & Buabeng-Andoh, 2022; Delen, 2010; Muncie, 2020; OECD, 2022a).

Several factors are mentioned in the existing literature as

contributors to the students' drop out or delayed to no graduation. To name but a few, while educators and teachers are said to be instrumental in facilitating the students learning processes (P. Mishra & Koehler, 2006; Ndukwe & Daniel, 2020; Okoye et al., 2020, 2022), on the other hand, there exist different factors that could impact the attrition or ineffective learning process for the students, e.g., perpetuating discrimination in education, academic achievement of the disadvantaged or vulnerable groups, gender bias and students perspectives about the teaching/learning process and outcome, adapting of the online teaching and distance learning methods, self-perception of university teachers about the use of ICT in the classroom, etc (Bjarnason & Thorarinsdottir, 2018; Buser et al., 2019; Ewing & Cooper, 2021; Fresen & Hendrikz, 2009; Guillén-Gámez et al., 2021; Jimoyiannis et al., 2020; Kafedžić et al., 2018; König et al., 2020; Mercader & Gairín, 2020; OECD, 2022b; Sun et al., 2019; Tzovla et al., 2021; UNESCO, 2020, 2022b). A recent report by the United Nations (UN) shows that in some regions, there may exist some certain level of bias in the assessment of the learning outcome of the students (UNESCO, 2020, 2023). This bias corresponds to a 4% difference in the likelihood of retention, and a 5% reduction in the probability of certain groups of students attaining the maximum grades or academic achievement in class (UNESCO, 2020). Parental education (academic achievement or qualification) including their income/socio-economic status has also been shown to greatly correlate with the choice or educational outcomes of the students (OECD, 2019; 2022a). It appears that the graduation rate of students who entered a degree program is disaggregated or related to the highest level of education attained by either of the parents (OECD, 2019). Although, across the countries, there is no concrete evidence or pattern between the parents' academic achievement and students who complete their program within the estimated theoretical duration.

2.2. Feature selection in prediction of students' retention and graduation

The questions "Where are students after their first year of study (retention)? and by the end of their program (graduation)?" (also used as the main features for prediction of the students' retention and graduation in this study) has proven to be one of the most pertinent ways to technically and pedagogically determine the effectiveness of the students' orientation (e.g., first time in higher education), management of the educational process or curriculum, and/or the outcome of education (OBE) (Amirtharaj et al., 2021; Chiu, 2020; OECD, 2019, 2022b; Okoye, 2022; UNESCO, 2015, 2020, 2022b). Available data shows that globally an average of 41% of tertiary education students are still enrolled, 20% no longer in any program, and 39% graduating within the theoretical duration (OECD, 2019; 2022a). Two timeframes or indicators are used to measure the students' retention and graduation, namely (i) the status of the student after their first year of study, and (ii) by the end of the theoretical duration of the program (Arqawi et al., 2022; Brdesee et al., 2022; Cabral Gouveia et al., 2023; Cardona & Cudney, 2019; Dake & Buabeng-Andoh, 2022; Delen, 2010; Malik, 2011; Muncie, 2020; Nayak et al., 2023; OECD, 2019, 2022a; Ploutz, 2018; Uliyan et al., 2021; Veenstra, 2009; Zhao et al., 2021). Based on the aforementioned timeframes, this present study has used the most widely applied feature in the literature corresponding to the students' retention or graduation, namely (i) cumulative grade point average (CGPA) of the student in their previous secondary or high school, and (ii) the entry exam or admission score, to train and build the ML model (RG-DMML) and algorithm described in this study. The selected features were used to predict the rate of the students' retention and graduation within the higher education setting. The analyzed features or variables (see Sections 3 and 4) prove to be effective by making use of the Wrapper method for feature selection that assesses the quality of machine learning approaches using different subsets of features against the evaluation criterion, and selecting the optimum set of features for which the model's performance is optimum (Maldonado et al., 2022).

2.3. Literature on ML approaches for prediction of students' retention and graduation

Machine learning (ML) is one of the fields in modern computing that is now widely used to predict and enable an improved experience for users across different domains or sectors (Bell, 2022; Shinde & Shah, 2018). In education, especially taking into account the countless diverse didactical areas of its application, ML methods have also been applied for predicting the rate of student retention and graduation (Arqawi et al., 2022; Brdesee et al., 2022; Cardona & Cudney, 2019; Dake & Buabeng-Andoh, 2022; Delen, 2010; Muncie, 2020; Nayak et al., 2023; Okoye et al., 2022; Palacios et al., 2021; Ploutz, 2018; Priyambada et al., 2023; Uliyan et al., 2021). Students' academic performance and outcome prediction using machine learning or AI methods and approaches, is one of the most important applications of educational data mining (EDM) that have proved to help to improve both the quality of the education process and delivery (Arqawi et al., 2022; Brdesee et al., 2022; Cardona & Cudney, 2019; Dake & Buabeng-Andoh, 2022; Delen, 2010; Muncie, 2020; Nayak et al., 2023; Okoye et al., 2022; Ploutz, 2018; Uliyan et al., 2021). The strengths and weaknesses of the different applied ML models and approaches may also depend on the application scenario or context in which the methods and algorithms are applied. As expounded in the methodology and analysis of this study (see Section 4.1), the ensemble machine learning algorithm is concerned with employing multiple base learners and combining their predictions to determine the overall performance of the implemented ML models or approach (LaViale, 2023; Mienye & Sun, 2022; Zhang et al., 2019). Thus, the multiple learners based on different categories of the training sets or values are combined and used to evaluate the overall performance accuracy of the implemented model and classification accuracy (LaViale, 2023; Mienye & Sun, 2022; Ngo et al., 2022; Sibindi et al., 2023).

More relevant studies and technical components as it concerns the use of AI/ML techniques for the prediction of students' retention and graduation are discussed here.

Dake and Buabeng-Andoh (2022) developed an ML algorithm to predict students' dropout rates and identify dominant attributes that affect the learners' attrition and retention. The model was deployed as an ensemble algorithm built and validated using different ML methods such as SVM (support vector machine), DT (decision tree), MLP (multilayer perceptron), and RF (random forest) algorithms. Delen (2010) proposed an ML approach that uses multi-layer perceptron (MLP) type artificial neural network architecture (ANN) to predict early student attrition or retention rate. The method based on the popular data mining methodology (CRISP-DM) as utilized in this study, uses historical data from student databases capable of predicting and explaining the institution-specific nature of the attrition problem. Similar but methodologically different from the work done in this study using ML technique, Uliyan et al. (2021) used the deep learning technique such as the condition random field (CRF) and bidirectional long short-term model (BLSTM) to study each students' label and those students whose retention was at risk. Their method and framework focused on the first-year students' retention and graduation rate for the undergraduates across the analyzed period. Ploutz (2018) also applied predictive modeling such as Logistic regression, Decision trees, SVM, and Neural networks to predict whether a student will graduate by analyzing the students' information related to admissions applications, financial aid, and first-year academic performance. Perhaps, based on the individual feature selection mechanism or feature ranking of the study (Ploutz, 2018), it suggests that educators emphasize core high school GPA and unweighted high school GPA as a more accurate measure of predicting success at higher institutions, such as the one done in this present study.

In other similar studies, Arqawi et al. (2022) predicted university students' retention, using data collected from Kaggle repository with a combination of supervised machine learning algorithms and deep learning (DL) method. Brdesee (2022) proposed an ML approach

capable of enhancing academic decision-making related to the students' performance. Their study reveals the effectiveness of using the long short-term memory (LSTM) technique for the prediction of at-risk students using a dataset containing demographics, learning/academic, and education-related attributes suitable for the development and deployment of the ML algorithms. Muncie (2020) using a combination of supervised ML methods such as Logistic regression, Decision tree, Gaussian naive Bayes, and the Administrative student data, designed quantitative research that focused on predicting whether a student would be retained following the semester after they enter a university program, taking into account set of features such as their weighted high school GPA. The students' data was used to classify them into risk categories to act as an early warning system for academic institutions. Okoye et al. (2022) proposed an educational process and data mining plus machine learning (EPDM + ML) model that contextually analyzes and automatically predicts what the students' recommendation scores for the teachers or learning assessment would be based on data from the student's evaluation of teaching (SET). Their method implemented using K-Nearest-Neighbor (KNN) and validated using *k*-fold cross-validation, uses the average sentiment and emotional valence (quantified) data about the students to predict the recommendation scores. Palacios et al. (2021) predicted student retention at each of three levels during their first, second, and third years of study in the higher education setting. Their method which was tested using the Decision tree, K-Nearest neighbors (KNN), Logistic regression, Naive bayes, Random forest, and Support vector machine (SVM) algorithms, all showed to adequately predict each of the levels when dropout occurs, and that secondary or previous educational score of the students are important predictive variables in such type of research, such as the one done in this study.

Technically, Mienye et al. (2022) surveyed the different types and applications of the ensemble machine learning approaches and algorithms in the literature, such as the Bagging, Boosting, and Stacking methods. On the one hand, Sibindi et al. (2023) highlighted the boosting technique (e.g., Adaboost, GBM, LGBM, XGBoost and catboost) as a method that has evolved throughout the years, particularly with the goal of improving the performance of the machine learning techniques and models. Recent studies have also emphasized on the Bagging method (LaViale, 2023; Ngo et al., 2022; Zhang et al., 2019) that allows for aggregation of the results from different versions of the same model in order to enhance the models' accuracy, as seen in this study (see Section 4, Tables 4 and 5). According to the study by LaViale (2023), the ensemble methods can improve the overall performance of ML models particularly the KNN models as introduced in this study, by combining the predictions of multiple versions of the model. The author (LaViale, 2023) emphasizes that one of the pertinent ways to do this is by using the Bagging technique, where multiple KNN models are trained on different subsets of the training data, and their predictions combined through averaging or voting. However, Ngo et al. (2022) also note that while "better performance" is the sole result of reduced variance using the Bagging method, on the other hand, a common challenge with such a method is that the bias of an individual machine learner is mostly likely to be the same as the bias of the combined models. Two-layer ensemble prediction framework has also been proposed to predict the students' performance based on the learning behavior and domain knowledge (Priyambada et al., 2023).

Table 2
Multiple Linear Regression Correlation matrix for the Retention vs Graduation data.

Co-linearity for CGPA and Admission score vs Retention and Graduation (Method = Pearson Correlation)				
Dataset		Retention status	CGPA	Admission score
Retention (n = 52262)	Retention status	1.000	.119	.076
	CGPA	.119	1.000	.296
	Admission score	.076	.296	1.000
		Graduation status	CGPA	Admission score
Graduation (n = 53639)	Graduation status	1.000	.204	−.104
	CGPA	.204	1.000	.170
	Admission score	−.104	.170	1.000

Note: Sig. (1-tailed) = 0.000.

In summary, based on our review of the literature and the work done in this study, we note that the ability to predict students' academic status or performance (e.g., the retention and graduation rate) is an important phenomenon not only for the student's success, but also for the educators in the different higher institutions of learning and settings. Automated predictions, suitable feature selection mechanisms, and improved models' performance that underlies the ML method (RG-DMML) introduced in this study, can be made and tested by using the (educational) datasets stored in information systems or databases of the various institutions about the students' behavior or record related to the individual courses or learning outcome and performance (Ngo et al., 2022; Okoye et al., 2022; Priyambada et al., 2023).

3. Methodology

This study followed the popular data mining and knowledge discovery methodology called Cross Industry Standard Process for Data Mining (CRISP-DM) (Delen, 2010; Martinez-Plumed et al., 2021; Peker & Kart, 2023; Schröer et al., 2021; Shearer, 2000; Wirth & Hipp, 2000). The CRISP-DM is a systematic and structured way of conducting data mining or machine learning studies by following six main steps as listed below:

1) Contextual knowledge and understanding of the domain and goals for the study.
2) Identification, assessment, and determination of the sources of data.
3) Pre-processing, cleaning, transforming, and normalization of the data.

Table 1
Descriptive Statistics and information about Selected features and Principal Component Factor Analysis (PCA).

Data	KMO	Bartlett's Test	p-value (sig.)	Selected Feature	Data type	Mean	Std. Dev.	Feature Description
Retention (***n*=52262**)	0.528	5627.88	0.00*	CGPA	num	89.00	6.31	average grade of student during previous high school
				Admission score	num	1344.54	187.06	results of the entering exam
				Retention status	num	0.910	0.287	target retention rate after 1st year, where 1:retention 0: no retention
Graduation (***n*=53639**)	0.424	4865.64	0.00*	CGPA	num	85.91	8.46	average grade of previous high school study
				Admission score	num	855.35	658.67	student grade after the entry exam
				Graduation status	num	0.730	0.444	if student finished study within expected period, where 0:no 1:yes

Table 3
Correlation Coefficient (ρ) results for the Retention vs Graduation dataset.

Model (ρ)		Unstandardized Beta (β)	Std. Error	Standardized Beta (β)	t	Sig.
Retention (n = 52262)	CGPA	.005	.000	.105	23.160	.000
	Admission score	.000	.000	.045	9.919	.000
Graduation (n = 53639)	CGPA	.012	.000	.228	53.801	.000
	Admission score	.000	.000	−.142	−33.557	.000

Note: Significant levels ($p \leq .005$), Response or Outcome Variable = Retention and Graduation status.

Table 4
Model performance and results for the *k*-values based on the 10-fold cross-validation.

k-fold cross-validation ($k = 10$)	Retention (n = 52262)		Graduation (n = 53639)	
Model 1	knn.224	100	knn.227	85.19
Model 2	knn.224	100	knn.228	85.22
Model 3	knn.226	100	knn.229	85.24
Model 4	knn.227	100	knn.230	85.32
Model 5	knn.228	100	knn.231	85.33
Model 6	knn.229	100	knn.232	85.35
Model 7	knn.230	100	knn.233	85.35
Model 8	knn.231	100	knn.234	85.38
Model 9	knn.232	100	knn.235	85.34
Model 10	knn.233	100	knn.236	85.39

Note: *k*-values computed using closest $k = \text{sqrt}(n)$, prediction rate at 0 for 0% and 100 for 100%.

Table 5
Result of performance and evaluation of the implemented model.

Model Performance	Retention	Graduation
Precision	0.909	0.822
Recall	1.000	0.957
Accuracy	0.909	0.817
Error-rate	0.090	0.182
F1 Score	0.952	0.885

4) Development of the predictive model using comparable analytical technique.
5) Evaluation and assessment of validity and efficacy of the implemented model in each run test or iteration of the model, and against the goals of the study.
6) Deployment of the model for use, e.g., practical application, implications for its real-time use, and decision-making.

A *k*-fold cross-validation (see Eqn. (1)) approach (Delen, 2010; Hastie & Tibshirani, 2009; Wong & Yeh, 2019; Xiong et al., 2020) was used to implement and assess the technical/practical impact of the proposed model to predict the students' retention and graduation rate or status. The proposed ML model (RG-DMML) and ensemble algorithm based the Bagging method (LaViale, 2023; Ngo et al., 2022; Zhang et al., 2019) for its application in real-time (see Fig. 1, Algorithm 1) was built and implemented using the K-Nearest-Neighbor (KNN) (Ali et al., 2023; Ghosh et al., 2020; Okoye et al., 2022; Viji et al., 2020). KNN methods consider neighboring objects in the analyzed datasets and have proven to train better machine learning models (Arqawi et al., 2022; Elzamly et al., 2015; Lubis et al., 2020).

This study shows that one of the best ways to implement the ML models, especially the KNN-based model described in Fig. 1, is through cross-validation. To select the cross-validation dataset from the training dataset, we take the small portion from the training dataset and call it a validation dataset, and then use the same to evaluate different data points or possible values of K (10-folds in this study) with equal weights (see Algorithm 1, Table 4). This way we can predict the label for each

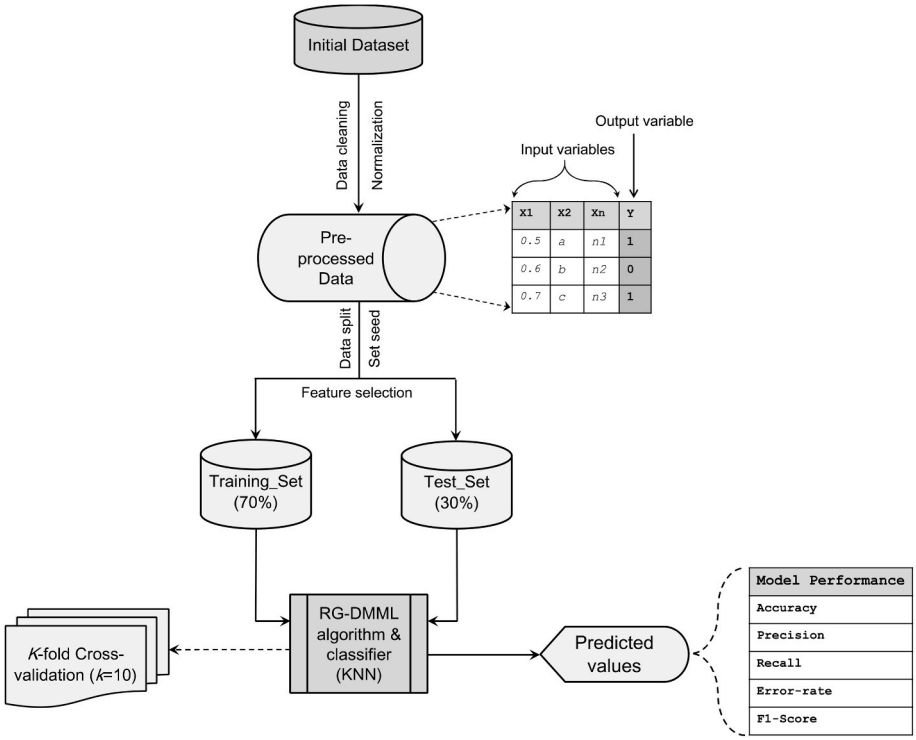

Fig. 1. Architectural overview of the proposed RG-DMML model.

version of the executed model or validation set where K equals to 1, K equals to 2, K equals to 3 etc ($k = 1, 2, 3 \ldots K_n$), and then consequently establish which value of K gives us the best performance on the validation set, therefore, minimizing the likelihoods of validation error.

Thus, as summarized in Fig. 1, the RG-DMML model is built and implemented based on the following definitions: if we assume that the implemented *k*-folds cross-validation or value of *k* is set from 1 to K_n, (for instance up to 10) which is a common practice in a lot of predictive data mining applications, then we explain in the following equations (see Eqn. (1) and 2) the method for estimating the optimal *k*-values as follows (Ali et al., 2023; Lubis et al., 2020);

For each $k = 1, 2, \ldots K_n$, we fit the model with parameter λ to other $K-1$ parts giving $\beta^{\wedge -k}(\lambda)$ which also helps to compute its error in predicting the *kth* part:

$$E_k(\lambda) = \Sigma i \in kth\ \text{part}(y_i - x_i\beta^{\wedge -k}(\lambda))^2 \quad \text{(Eqn1)}$$

The implementation of the above process (Eqn (1)) results in cross-validation checking of value:

$$CV\ (\lambda) = \frac{1}{K}\sum_{k=1}^{k} E_k\ (\lambda) \quad \text{(Eqn2)}$$

The steps (Eqn. (1) and Eqn. (2)) are repeated for many values of λ, and then the value of λ that produces the smallest *CV* (λ) is chosen, as described and explained in the proposed algorithm (see Algorithm 1) and executed *k*-folds cross-validation (see Fig. 2).

For the classification process using the KNN model or classifier (see Algorithm 1); the output of the method is defined and computed as an object assigned to the class that is most common among the *k*-nearest neighbors (see Eqn. (3) and 4). This means that if we have a dataset with labels *x* and *y* (see illustration in Fig. 1) and want to predict the link between the variables. Then our goal is to discover the function h:X→Y wherefore having an unknown observation *x*, h(*x*) we can positively predict the identical output *y* (see Fig. 1 and Algorithm 1).

To determine the distance metrics, the Euclidean metric is defined as follows:

$$d(x, x') = \sqrt{(x_1 - x_1')^2 + \ldots + (x_n - x_n')^2} \quad \text{(Eqn3)}$$

Finally, the input variable(s) *x* is assigned to the class with the largest probability defined as follows:

$$P(y = j \mid X = x) = \frac{1}{K}\sum_{i \in A} I\left(y^{(i)} = j\right) \quad \text{(Eqn4)}$$

3.1. Data description

Two sets of data were used in the data analysis. Data was collected and consolidated from the university student database at a private university in Mexico. The dataset consisted of information about the (i) Retention, and (ii) Graduation status of the undergraduates collected over the past 10 years at the institution where this study was conducted. For ethical purposes, we note that the data was anonymized upon gathering and use for this study. The average students' retention rate for the institution is about 92%, and the average rate for graduation is about 75% for the analyzed period.

The datasets after assessing and preprocessing the files to identify and remove anomalies/unusable records consisted of a total sample of $n = 52262$ for the Retention, and $n = 53639$ for Graduation. The data comprised of different variables related to the students' demographic information, academic record, and retention/graduation status, such as year of entry, previous academic record, campus, type of program, admission score, school, enrollment status, gender, age, scholarship, etc. However, in Table 1 the authors provide only the list of the analyzed features (variables) it has selected for the study based on the Wrapper feature selection method (Maldonado et al., 2022). This includes the most commonly used feature in the literature that has proven to be most effective for training and testing the measured constructs (retention and graduation); thus, (i) the average grade of the previous high school (CGPA), and (ii) the entry/admission score of the students (OECD, 2019; 2022a).

As presented in Table 1, the Principal Component Factor Analysis (PCA) with Varimax Rotation where Eigenvalue>1 (Allen, 2017; Brown, 2019; Frost, 2020) was used to test the selected features or variable, and the results of the analysis shows that the individual features where adequate and reliable for testing both constructs (retention and graduation); with Kaiser–Meyer–Olkin Measure of Sampling Adequacy (KMO) = 0.528, Bartlett's Test of Sphericity = 5627.88, and p-value (sig.) = 0.00 for the Retention data, and KMO = 0.424, Bartlett's Test of Sphericity = 4865.64, and p-value (sig.) = 0.00 for Graduation data, respectively (Ermatita et al., 2019). Also, for the descriptive statistics, we note that there is no significant difference between the genders of the students in the data with 55% records for males and 45% for females for Retention data, and 57% for males and 43% for females for Graduation data, respectively.

In Table 1, the authors present the Descriptive statistics for the analyzed data which includes the mean distribution of the CGPA scores and entrance examination scores for the retention and graduation datasets, respectively. In Tables 2 and 3 we present the correlation matrix (Table 2) and correlation coefficient (ρ) (Table 3) results for the individual features or variables in the data which was done using Multiple Linear Regression analysis (Olive, 2017). The multiple co-linearity results and correlation coefficient (ρ) reported in the tables (see Tables 2 and 3) shows that the explanatory or predictor variables (i.e., CGPA and Admission scores) are linearly related or are said to be a good (effective) feature for predicting the outcome of the response variables (i.e., Retention and Graduation status), respectively. With the Standardized Beta (β) value for the CGPA (Retention = 0.105, Graduation = 0.228) and Admission score (Retention = 0.045, Graduation = −0.142) for the two datasets (retention and graduation) showing to be significant ($p = .000$), where the significant level is statistically measured at $p \leq .05$, respectively.

Furthermore, from the results presented above (Tables 2 and 3), it is noteworthy to emphasize the fact that while both variables, i.e. the CGPA score and the Entrance/Admission score, proved to be significant (p = .000) and are important features in predicting the two factors (Retention and Graduation status). On the other hand, it can be said that the CGPA tends to be a more effective feature in predicting the two

```
# Computing the predictions for each value of the k-fold (10), where approximate square root k=sqrt(n), k = 224 to 233, Number of observations n=52262
i=1
k.optm=1
for(i in 224:233) {
  knn.model <- knn(train = train.RG_MachineLearning, test = test.RG_MachineLearning, cl=train.RG_MachineLearning_labels$Retencion, k=i, prob = TRUE, use.all = TRUE)
  knn.model <- as.numeric(levels(knn.model))[knn.model]
  k.optm[i] <- 100 * sum(test.RG_MachineLearning_labels = knn.model) / NROW(test.RG_MachineLearning_labels)
  k=i
  cat(k, "=" ,k.optm[i], "\n")
}
```

Fig. 2. Fragment of the implemented algorithm for computing the optimal *k*-values using *k*-folds cross-validation method.

constructs (retention and graduation status) than the admission score (see Tables 2 and 3), vice and versa.

4. Data analysis and model implementation

Using the CRISP-DM methodology for implementation of the defined model (RG-DMML) (see Methodology – Section 3, Fig. 1); the study applied the following algorithm (Algorithm 1) for the data analysis process and results extraction.

Algorithm 1. Ensemble algorithm for prediction of students' retention and graduation

```
1: Input:      Initial dataset
2: Output:     Confusion matrix, Model Performance: precision, recall, accuracy, error-rate, F1-score
3: Procedure: ML model and classifier: Predictive model
4: Begin
5:   For all extracted dataset, RG
6:   Select input variables and features
7:     while no more features or considerations (A) is left do
8:   Create machine learning subset data, ML, by concatenating c(selected features) ← from RG
9:   Normalize object ML ← function(x) = return((x - min(x)) / (max(x) - min(x)))
10:  Set seed, Training (TR) and Test (TS) dataframes ← from ML
11:  Create separate dataframes for predicted variable results, TR_labels and TS_labels ← from RG
12:  Calculate estimated correct classification for k-folds and predicted scores for TR and TS
       If TR and TS interpretation ← Null then
         obtain the relevant features A from RG and loop to line 8
       Else If TR and TS interpretation ← 1 then
         Run knn( ) method
           i=1
           k.optm=1
           for (i in k_range1 : k_range2) {
                knn.model ← knn(train = TR, test = TS, cl= TR_labels$PredictionFeature, k=i)
                k.optm[i] ← 100 * sum(TS_labels = knn.model) / NROW(TS_labels)
                k=i
                cat(k, "=" , k.optm[i], "\n")
               }
13: Return: outputs, prediction levels for all k-folds in all knn.model and iterations
14: End If statements
15: End while
16: End For
17: End
```

As gathered in Algorithm 1, the RG-DMML model was deployed using knn( ) method in R integrated development environment and programming software (Rstudio, 2023); whereby the following method was used for the execution process: knn(train = data, test = data, cl = train_labels$PredictionFeature, k = i) (see Line 12). The process involved the successive application of the different steps we listed earlier by following the CRISP-DM methodology (Martinez-Plumed et al., 2021; Peker & Kart, 2023; Schröer et al., 2021; Shearer, 2000; Wirth & Hipp, 2000) (see Section 3). The Input module were initial data collected about the retention and graduation status of the students (see Data description - Section 3.1). The Output were confusion matrix table used to calculate the model performance, i.e., Precision, Recall, Accuracy, Error-rate, and F1-score (see Table 5 and Fig. 1).

To describe the prediction and classification process; we extracted and selected the input variables or features based on the Wrapper selection method (Maldonado et al., 2022), where in turn, the students' CGPA and Admission scores was used as the predictor variables, and the Retention and Graduation status used as the response variables, respectively (see Tables 1–3) (Algorithm 1, Lines 5 and 6). It is important to mention that the selected features were concatenated and normalized (Lines 8 and 9) before the process of creation of the dataframe and splitting of the datasets (i.e., training set = 70%, test set = 30%) (Lines 10 and 11), which were used in estimating the performance/accuracy of the models' classification and predicted scores done using the knn( ) method and validated through the *k*-folds (10-fold) cross-validation (Lines 12 and 13).

It is important to mention that the implemented model and ensemble algorithm or process for calculating the estimated correct classification for the 10-folds cross-validations and results, as reported in Table 4 and Fig. 2, was done by computing the predictions for each *k*-values whereby the approximate square root is equal to k = sqrt(n). Therefore, for each number of observations for the Retention (n = 52262) and Graduation (n = 53639) datasets, we used the closest Sqrt(n), where n is the total number of data points that most helps in selecting a better model, thus, knn = 224 to 233 for Retention, and knn = 227 to 236 for Graduation, respectively.

Fig. 2 is an example of the implemented algorithm and executed code for computing the optimal *k*-values or model performance for each of the 10-fold run tests, illustrated using the retention dataset. Fig. 3(a) and (b) and Table 4 show the results of the different knn models and iterations including the performance of each version of the model.

4.1. Model performance and evaluation

In the confusion matrix statistics for the implemented models (see Fig. 4(a) and (b)), our careful analysis of the output and results shows that the prediction rate or accuracy for the "Yes" class (i.e., 1) were significantly higher than the prediction rate or accuracy of the "No" class (i.e., 0). This was observed for both sets of data (retention and graduation), although the model showed to be more efficient in predicting the students' retention (Fig. 3(a) and 4(a)) in comparison to the students' graduation data (Fig. 3(b) and 4(b)).

Based on the confusion matrix (computed using the hold-out samples

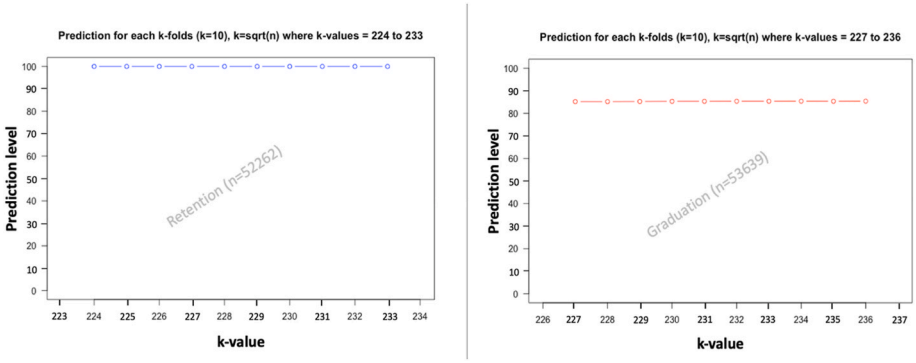

Fig. 3. (a) Prediction rate for Retention dataset.
Fig. 3(b)Prediction rate for Graduation dataset.

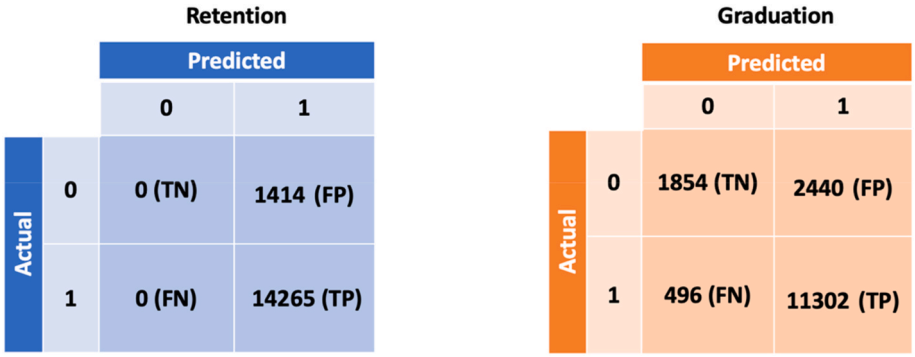

Fig. 4. (a) Performance metrics for Retention data.
Fig. 4(b)Performance metrics for Graduation data.

for testing and validation of the model, i.e., $n = 15679$ data points for Retention, and $n = 16092$ for Graduation, that constitute of 30% of the original dataset selected and used as the test set) (see Fig. 1 and Algorithm 1), we evaluated the performance of the implemented model by using the following evaluation metrics (A. Mishra, 2018; Muntean & Militaru, 2023; Nayak et al., 2023; Okoye et al., 2022; van der Aalst, 2016):

- TP—number of true positives, representing instances of the scores that were correctly classified as positive.
- TN—number of true negatives, representing instances of the scores that were correctly classified as negative.
- FP—number of false positives, representing instances of the scores that were predicted to be positive but should have been classified as negative.
- FN—number of false negatives, representing instances of the scores that were predicted to be negative but should have been classified as positive.

Therefore, using the confusion matrix statistics or classification results of the model (Fig. 4(a) and (b)), both for the retention and graduation data; we calculated the performance of the implemented model by determining the Precision, Recall, Accuracy, Error-rate, and F1-scores as follows:

- Precision which refers to the number of positive predicted values is calculated as follows: (TP)/(TP + FP)
- Recall which refers to the true positive rate is calculated as follows: (TP)/(TP + FN)
- Accuracy of the model is calculated as follows: (TP + TN)/(TP + TN + FP + FN)
- Error-rate which refers to how often the model classification is wrong is calculated as: (FP + FN)/(TP + TN + FP + FN)
- F1-score which refers to the harmonic mean of the precision and recall is calculated as follows: (2 × Precision x Recall)/(Precision + Recall)

The results of the model performance and evaluation for the retention and graduation data are presented in the following table (Table 5).

5. Discussion

In the confusion matrix of the implemented RG-DMML model (Fig. 4 (a) and (b)), the outcome shows that the prediction accuracy for the "Yes" class (i.e., 1) for the two datasets (retention and graduation) were significantly higher than the prediction rate for the "No" class (i.e., 0). This was also evidenced in the performance evaluation of the model (see Fig. 3(a) and (b), Tables 4 and 5) therein we found for both cases, that the implemented model proved to be efficient, with a high level of performance and accuracy in predicting the students' retention and graduation status. Thus, with Precision (Retention = 0.909, Graduation = 0.822), Recall (Retention = 1.000, Graduation = 0.957), Accuracy (Retention = 0.909, Graduation = 0.817), F1 score (Retention = 0.952, Graduation = 0.885), and low error-rate (Retention = 0.090, Graduation = 0.182), respectively.

Considering the analyzed features used in the prediction of the students' retention and graduation status (i.e., the previous academic CGPA and Admission scores); it can be said that the implemented ML model proved effective in predicting both the retention and graduation status of the students (see Table 4). It is important to also mention from the results presented above that the model proved to be more efficient in the prediction of the students' retention status or dataset (Fig. 3(a) and 4(a)) in comparison to the students' graduation data (Fig. 3(b) and 4(b)). This

very observation (see Fig. 3(a) and (b), 4(a) and 4(b)) suggests that machine learning approaches or models that tend to use those features (i.e., previous academic CGPA and Admission scores of the students) in modelling or prediction of the retention and graduation status of the students, are more apt to effectively predict the learners who are likely to drop out of their program (retention status) with a higher level of accuracy, than those who are predicted to complete the program (graduation status), vice and versa.

Furthermore, our review of the existing literature shows that there are not too many studies that have looked into the graduation rate of the students (Ploutz, 2018; Uliyan et al., 2021), such as the one done in this study, in comparison to the much greater studies that have investigated the retention or drop-out rate of the students in higher education (Arqawi et al., 2022; Cardona & Cudney, 2019; Delen, 2010; Muncie, 2020; Palacios et al., 2021). We note that future research studies can consider looking into this promising areas or gap highlighted in this study. This will help not only in generalizing the results of this present study, but also in its mechanism in terms of transferability to include other components or factors that may help in identifying the common conundrum or intersection between the retention and graduation rate or status of the students in higher education. This underlines the technical implications, impact, and importance of the feature selection mechanisms in the prediction of the students' retention and graduation in education, which is one of the key contributions of this study. Interestingly, existing studies have also highlighted the didactical and technical implications of the students' retention and graduation rate modeling in higher education primarily determined by the attributes (features) used for the automated prediction or algorithms development (Dake & Buabeng-Andoh, 2022; Delen, 2010). Likewise, similar studies that focus on the *predictive modeling techniques* have also pointed out the sensitivity of the resultant (predictive) models, by stating that the most important predictors for the students' attrition are those features (or parameters) that are related to the past and present educational success of the student (Delen, 2010), such as previous academic grades (CGPA) and entrance (admission) examination scores of the students used in this study (see Tables 1 and 2).

Keeping this view in mind, in our review of the available literature and triangulation to the above observations or mechanism, it shows that two timeframes (indicator) are used to measure the students' retention or graduation rate; (i) the status of the student after their first year of study, and (ii) by the end of the theoretical duration of the program (Arqawi et al., 2022; Brdesee et al., 2022; Cabral Gouveia et al., 2023; Cardona & Cudney, 2019; Dake & Buabeng-Andoh, 2022; Delen, 2010; Malik, 2011; Muncie, 2020; Nayak et al., 2023; OECD, 2019, 2022a; Ploutz, 2018; Uliyan et al., 2021; Veenstra, 2009; Zhao et al., 2021). In this vein and perspective, this study has used the most relevant and widely applied feature in the available literature, namely (i) CGPA of the students in their previous study, and (ii) the entry exam or admission score (see Table 1) (Arqawi et al., 2022; Brdesee et al., 2022; Delen, 2010; OECD, 2022a; Palacios et al., 2021; Ploutz, 2018), to train and test the machine learning model (RG-DMML) and ensemble algorithm described in this study (Fig. 1, Algorithm 1). The output of the model and outcome proved to be efficient and effective for the prediction of the students' retention and graduation status with a high level of accuracy and model performance (see Section 4).

However, taking into consideration the generalisability of the findings of this study, the authors note that not all university admissions systems take into account the CGPA and entrance exams, and the fact that these same features (CGPA and entrance examinations) are used by several other universities to determine offers to the courses. For example, in situations whereby a CGPA cut-off is applied and only students with CGPA above a certain threshold are offered entrance examinations. It is important to note that the multiple co-linearity test we conducted to determine whether the two variables (CGPA and entrance examinations) are correlated given the retention and graduation rate or status of the students (see Section 3.1, Tables 2 and 3), shows that both features (CGPA and entrance examinations) are correlated and are good predictors of the measured constructs (retention and graduation) across the data. Therefore, administrative and pedagogical -wise, it suggests that educators or university admissions systems should take into account the CGPA and entrance exams of the students upon recruitment especially in support of the different data management strategies and governance of the students' learning process and progression, curriculum design and learning outcomes, or the prediction of the students' retention and graduation status, as done in this study. Besides, the scientific evidence we drew from the existing literature shows that the two features (CGPA and entrance examinations) are the most widely used in the assessment or prediction of students' attrition and dropout rate in education (Brdesee et al., 2022; Dake & Buabeng-Andoh, 2022; Delen, 2010). Indeed, another factor that could also explain the implications or outcomes of the machine learning procedure and work done in this study is whether university curriculums are similar to the later years of high-school of the students. For instance, existing research has looked into statistics about how high-school grades and university performance can serve as an input into estimating the subsequent performance of first year students (Cyrenne & Chan, 2012). A recent study on predicting students' academic performance at secondary and intermediate level using machine learning techniques and classifiers (Hussain & Khan, 2023), shows that the AI/ML technologies are efficient and relevant for predicting the students' performance or academic status as shown in the results and outputs of this study (see Section 4). Moreover, by tracking the university performance of the students, studies have been able to estimate the likelihood of success of subsequent students or curriculum based on their characteristics as well as their high school grades (Arqawi et al., 2022; Brdesee et al., 2022; Cyrenne & Chan, 2012; Dake & Buabeng-Andoh, 2022; Hussain & Khan, 2023).

5.1. Implications of the study

One of the main implications of this study is that educators and many institutions can effectively adopt the AI/ML-based or yet advanced analytical methods (by making use of the available information or data in the databases) to accurately predict and monitor those learners who are likely to drop out of their programs or delayed to no graduation. Therefore, optimizing/heightening the outcome of the learning process and progression, institutional resources, and curriculum development for the stakeholders (students, teachers, educators), etc (Cabral Gouveia et al., 2023; Dake & Buabeng-Andoh, 2022; Delen, 2010; OECD, 2022a, 2022b). It is noteworthy to mention, that whilst there are research studies that have investigated the topic by studying the use of AI/ML methods to detect the retention or drop-out rate of students (Arqawi et al., 2022; Brdesee et al., 2022; Cardona & Cudney, 2019; Delen, 2010; Muncie, 2020; Palacios et al., 2021), and only a few studies that have researched the graduation rate of the students (Ploutz, 2018). We note, on the other hand, that this study is one of the only study to the best of our knowledge that looked into the use of the AI/ML technique for the prediction of both constructs (i.e., retention and graduation status of the students in higher education). Except for the study by Uliyan et al. (2021) which has only used the deep learning technique to study the students' label and those students whose retention and graduation may be at risk.

Technically and pedagogically -wise, the impact or implications of this study can be described to focus on three main aspects or context, primarily highlighted in the research questions of the study:

- Building and development of predictive/advanced analytical methods for educational process/data management through the implementation of machine learning methods or classification algorithms with a high level of performance and accuracy.
- The use of feature selection mechanism to identify key features or parameters that are most effective for modeling and development of

the AI/ML-based methods for prediction of students' retention and graduation status in education.
- Understanding of leading factors and impact of the AI systems and applications in identifying at-risk students for timely intervention, counselling, or decision-making, especially toward enabling a sustainable educational practice by and for the educators.

In the wider spectrum of scientific research and governance; we note that students' dropout, delayed completion, or even no completion can be costly to not only the students themselves, but to both the educators, governments, and the immediate family. Current data has shown that persons with higher education degrees are more likely to earn higher and contribute to the society (OECD, 2019; 2022a). Therefore, dropping out or delayed graduation may mean students or governments not reaping the full benefits of the degree or until upon completion. In this light, this study strongly believes that the theoretical and practical understanding of where students are after their first year of study (retention) and by the end of their individual programs (graduation) can be one of the promising and effective ways of managing the so-called challenges and gaps per se. This is explicated or analyzed in this study using the relevant features, such as the average grade of the student during their previous high school and results of the entry or admission exam (OECD, 2022a), to develop the ML model and algorithm for prediction of the students' retention and graduation. Consequentially, this highlights the much-needed technical requirement for forecasting and accurate prediction of non-trivial factors behind students' retention and graduation, particularly in higher education. Research studies such as this one can help educators and authorities in shaping up the educational systems/processes or the formative guidelines, e.g., by automatically identifying at-risk students or potential students' failure, and intervening in a timely manner.

Furthermore, while this study has shown that the capability to automatically predict the student's retention and graduation status in education using the student's data, can enable the many higher institutions of learning to take preventive measures to avoid learners' dropout or delayed to no graduation. On the other hand, other factors may serve as a challenge to its perceived ease-of-use and/or institutional effectiveness. For example, whereas the study has used two sets of data about the students' retention and graduation, and the widely known features for testing the two constructs (retention and graduation) such as the previous academic history of the student, and the entry exam status (see Sections 3.1 and 4). The authors note that scientific evidence has also shown that delayed graduation or dropout of students may not necessarily indicate the immediate student or institutional failure, but may also be linked to other external factors such as perpetuation of discrimination in education, academic attainment of the disadvantaged or vulnerable groups, parental education or socio-economic status, issues of gender bias, online teaching or self-perceiveness of the instructors about the use of ICT or educational technologies in the classroom (OECD, 2019, 2022a, 2022b; UNESCO, 2020, 2022b), which are highlighted to some degree in our review of the literature but are not discussed in full detail or as part of the methodology in this study. Future studies can investigate these further topic areas, especially from a data-driven or AI/ML-based sustainable educational practice or goal that aim to promote the quality of educational outcomes or SDG4 (Global Goals, 2022; UNESCO, 2023).

Institutional and country-specific factors have also been shown to explain the disparities or delayed to no graduation across the different countries or economies. For instance, it can be common in some countries for students to undertake a foundation or remedial courses that are not usually included as part of the formal curriculum or degree programs (OECD, 2019; Zhao et al., 2021), and disparities in graduation rates between the lengthier and shorter periods does not supposedly mean negative learning outcomes (Amirtharaj et al., 2021; Mukesh S, 2022). The higher institutions of learning are now beginning to adopt a flexible system in the offered programs, accentuated by the recent global pandemic (Shambour & Abu-Hashem, 2022; UNESCO, 2021), whereby, including in full-time programs, the learning contents or outcomes are divided into certain amount of credits that the students need to acquire in order to graduate, and not necessarily divided into number of years it takes the student to acquire the credits (OECD, 2019). Thus, on the one hand, while the flexible educational systems have proven to be beneficial for the students' learning and outcomes in the different hemispheres and dimensions (Müller & Mildenberger, 2021; Okoye et al., 2021). On the other hand, there still exist plausible tendencies of it increasing the number of students who do not graduate on time or within the estimated duration of study (OECD, 2019). Therefore, researching further into the impact of the socio-technical and educational/cultural factors listed earlier in this section (e.g., parental education or socio-economic status, issues of gender bias, online teaching or self-perceiveness of the instructors about the use of ICT in the classroom) especially as it concerns the students' retention and graduation in education may also serve as a way of complimenting the results and findings of this study.

Finally, in triangulation of the study's outcomes toward its implication for effective prediction of students' retention and graduation done in this study (Section 4), and by considering the high average students' retention (92%) and graduation (75%) rate for the analyzed period reported for the host institution where this study was conducted (see Section 3.1). We note that these results can be explained to link to some pedagogical or didactical factors such as the teaching approach employed by the different higher institutions of learning. For example, the TEC21 educational model (TEC, 2023a) applied by the host institution involves the design of study plans or curriculum that incorporate as a central unit of learning - the challenge-based education or learning (CBL) (Leijon et al., 2022; TEC, 2023b; Torres-Barreto et al., 2020) and the flexible digital learning components (MFD) (TEC, 2020a; 2020b) in which students develop disciplinary and transversal skills by solving challenges linked to real-world problems and demonstrating their mastery through various evidence of learning. The teaching approach (TEC, 2023a) encourages teachers to create active learning environments and find challenges that trigger the formation of disciplinary and transversal skills that the students require, as well as guiding them to transfer that knowledge to real contexts or work in a multidisciplinary way. Moreover, another didactical factor that can be idiosyncratically linked to the improved outcomes observed in this study is the students' opinion survey (ECOA, 2013), an instrument applied by the host institution to collect information about the teaching and learning performances of the students during and at the end of their respective courses. During the learning process or curriculum, the students are intermittently asked to complete the teaching evaluations or questionnaire in quartiles (after 6 weeks, 12 weeks, 18 weeks, and a final evaluation at the end of the courses) covering both during- and -end of their course. This helps the institution to keep track of the students' learning progression and to identify potential challenges that need to be addressed during the learning process, and in turn, used to ensure an effective learning outcome for the students or curriculum/policy management. Therefore, not only does this prove to increase or heighten the retention and graduation rate of the students, but also can serve as an effective tool or method to deal with the many challenges and barriers related to the students' dropout or delayed to no graduation across the many higher institutions of learning (Arqawi et al., 2022; Brdesee et al., 2022; Cardona & Cudney, 2019; Delen, 2010; Muncie, 2020; Palacios et al., 2021; Uliyan et al., 2021).

5.2. Limitations and future work

The ML model (RG-DMML) and ensemble algorithm through the Bagging method proposed in this study uses the 10-fold cross-validation and K-Nearest-Neighbor (KNN) supervised learning that technically considers neighboring objects in the data, and have proven to train better machine learning models. However, we note that the proposed model can further be transferred or generalized to other studies that may

consider using a different number of folds, different machine learning algorithms, or different feature selection techniques to study the constructs (retention and graduation), which could potentially compliment or produce different results. For instance, future studies can utilize the methodology adopted in this paper (i.e., RG-DMML model based on the CRISP-DM methodology) to study or predict the retention and graduation status of the students in different educational settings, or yet use different machine learning methods or feature selection techniques whilst at the same time guaranteeing the predictive ability and accuracy of the proposed model.

In addition, even though the ensemble methods or approaches have shown to provide excellent results or outcomes for solving machine learning problems, e.g., by allowing for aggregation of the results/outputs from different versions of the same model (e.g., one way to do this is by using the Bagging technique, where multiple KNN models are trained on different subsets of the training data and their predictions combined through averaging or voting) (LaViale, 2023), as shown in this study, and has proved to enhance the overall models' accuracy (see Section 4). However, it is also important to mention that the ensemble methods can come with a few drawbacks. For instance, using multiple versions of models for training in ensemble techniques could mean spending more computational time and resources. Moreover, whereas adding new models to an ensemble method purportedly means better predictions or performance, on the other hand, it may also make it cumbersome to trace the logic behind the decisions made by the AI algorithm. Besides, a common challenge with the Bagging method in machine learning is that the bias of a particular individual learner or model is likely to be similar to the bias of the combined model, as better performance is the sole result of reduced variance in such type of models (Ngo et al., 2022).

6. Conclusion

This study proposed an ML model (RG-DMML) and ensemble algorithm for the prediction of the students' retention and graduation status in education based on the CRISP-DM methodology. The model through the Bagging method and Wrapper feature selection method was implemented using the supervised machine learning technique - K-Nearest Neighbor (KNN), and validated using *k*-fold cross-validation. The results show that the executed model and ensemble algorithm was efficient and effective in predicting the student's retention and graduation status, with high performance and accuracy levels, whereby Precision (retention = 0.909, graduation = 0.822), Recall (retention = 1.000, graduation = 0.957), Accuracy (retention = 0.909, graduation = 0.817), F1-score (retention = 0.952, graduation = 0.885), and low Error-rate (retention = 0.090, graduation = 0.182), respectively. By considering the individual features utilized in predicting the model outputs, such as the (i) average grade of the previous high school, and (ii) entry/admission score of the students, the implemented model proved more efficient for predicting the students' retention status compared to the graduation data. In addition, the study empirically sheds light on the implications of the model outputs and factors that may impact the effective adoption or identifying of at-risk students for timely intervention, counselling, or decision-making toward a sustainable educational practice at large.

Funding

This research was supported by the host institution.

Statement on open data and ethics

The datasets used in this study are not publicly available and have been retrieved from the host institutional effectiveness management on request. All materials (dataset and associated codes and scripts) are available from the corresponding author on sufficient request.

CRediT authorship contribution statement

Kingsley Okoye: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Julius T. Nganji:** Conceptualization, Methodology, Validation, Visualization, Writing – review & editing. **Jose Escamilla:** Funding acquisition, Resources, Supervision, Validation, Writing – review & editing. **Samira Hosseini:** Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Validation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The authors would like to acknowledge the technical and financial support of the Writing Lab, Institute for Future of Education, Tecnologico de Monterrey, in the publication of this work. We will also like to acknowledge the Institutional department for retention and graduation, Tecnologico de Monterrey, for provision of the dataset used for the analysis in this study.

References

Ali, A. H., Mohammed, M. A., Hasan, R. A., Abbod, M. N., Ahmed, M. S., & Sutikno, T. (2023). Big data classification based on improved parallel k-nearest neighbor. *TELKOMNIKA (Telecommunication Computing Electronics and Control), 21*(1), 235–246. https://doi.org/10.12928/TELKOMNIKA.V21I1.24290

Allen, M. (2017). Factor analysis: Varimax Rotation. In *The SAGE encyclopedia of communication research methods*. SAGE Publications, Inc. https://doi.org/10.4135/9781483381411.n191.

Amirtharaj, S., Chandrasekaran, G., Thirumoorthy, K., & Muneeswaran, K. (2021). A systematic approach for assessment of attainment in outcome-based education. *Higher Education for the Future, 9*(1), 8–29. https://doi.org/10.1177/23476311211017744

Arqawi, S. M., Zitawi, E. A., Rabaya, A. H., Abunasser, B. S., & Abu-Naser, S. S. (2022). Predicting university student retention using artificial intelligence. *International Journal of Advanced Computer Science and Applications, 13*(9), 315–324.

Bell, J. (2022). What is machine learning? *Machine Learning and the City*, 207–216. https://doi.org/10.1002/9781119815075.CH18

Bjarnason, T., & Thorarinsdottir, B. (2018). The effects of regional and distance education on the supply of qualified teachers in rural Iceland. *Sociologia Ruralis, 58*(4), 786–804. https://doi.org/10.1111/SORU.12185

Brdesee, H. S., Alsaggaf, W., Aljohani, N., & Hassan, S.-U. (2022). Predictive model using a machine learning approach for enhancing the retention rate of students at-risk. *International Journal on Semantic Web and Information Systems, 18*(1), 1–21. https://doi.org/10.4018/IJSWIS.299859

Brown, J. D. (2019). Principal components analysis and exploratory factor analysis - definitions, differences, and choices. *Shiken: JALT Testing & Evaluation SIG Newsletter, 13*(1), 26–30.

Buser, W., Hayter, J., & Marshall, E. C. (2019). Gender bias and temporal effects in standard evaluations of teaching. *AEA Papers and Proceedings, 109*, 261–265. https://doi.org/10.1257/PANDP.20191104

Cabral Gouveia, M. D. C., Menezes, I., & Neves, T. (2023). Educational strategies to reduce the achievement gap: A systematic review. *Frontiers in Education, 8*. https://doi.org/10.3389/FEDUC.2023.1155741

Cardona, T. A., & Cudney, E. A. (2019). Predicting student retention using support vector machines. *Procedia Manufacturing, 39*, 1827–1833. https://doi.org/10.1016/J.PROMFG.2020.01.256

Chiu, M.-S. (2020). Exploring models for increasing the effects of school information and communication technology use on learning outcomes through outside-school use and socioeconomic status mediation: The ecological techno-process. *Educational Technology Research & Development, 68*, 413–436. https://doi.org/10.1007/s11423-019-09707-x

Cyrenne, P., & Chan, A. (2012). High school grades and university performance: A case study. *Economics of Education Review, 31*(5), 524–542. https://doi.org/10.1016/j.econedurev.2012.03.005

Dake, D. K., & Buabeng-Andoh, C. (2022). Using machine learning techniques to predict learner drop-out rate in higher educational institutions, 2022 *Mobile Information Systems*. https://doi.org/10.1155/2022/2670562.

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems, 49*(4), 498–506. https://doi.org/10.1016/J.DSS.2010.06.003

ECOA. (2013). *Student Opinion Survey (ECOA) - (Encuesta de opinión de los alumnus)*. Available at: https://portalrep.itesm.mx/va/encuestas/1.htm. (Accessed 15 July 2020).

Elzamly, A., Hussin, B., Naser, S., Abu, D. S., & Mohamed. (2015). *Classification of software risks with discriminant analysis techniques in software planning development process*. https://philpapers.org/rec/ELZCOS.

Ermatita, Isnainiyah, I. N., Yulnelly, Y., & Balqis, A. N. (2019). Usability analysis using principal component analysis (PCA) method for online fish auction application. In *Proceedings - 1st international conference on informatics, multimedia, cyber and information system* (pp. 231–236). ICIMCIS. https://doi.org/10.1109/ICIMCIS48181.2019.8985225, 2019.

Ewing, L. A., & Cooper, H. B. (2021). Technology-enabled remote learning during covid-19: Perspectives of Australian teachers, students and parents. *Technology, Pedagogy and Education, 30*(1), 41–57. https://doi.org/10.1080/1475939X.2020.1868562

Fresen, J. W., & Hendrikz, J. (2009). Designing to promote access, quality, and student support in an advanced certificate programme for rural teachers in South Africa. *International Review of Research in Open and Distance Learning, 10*(4). https://doi.org/10.19173/IRRODL.V10I4.631

Frost, H. R. (2020). *Eigenvectors from eigenvalues sparse principal component analysis (EESPCA)*. Available at: http://arxiv.org/abs/2006.01924. (Accessed 21 August 2020).

Ghosh, C., Saha, S., Saha, S., Ghosh, N., Singha, K., Banerjee, A., & Majumder, S. (2020). Machine learning based supplementary prediction system using K nearest neighbour algorithm. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3517197

Global Goals. (2022). *Goal 4: Quality education - the global goals*. Available at: https://www.globalgoals.org/goals/4-quality-education/. (Accessed 17 March 2022).

Guillén-Gámez, F. D., Mayorga-Fernández, M. J., & Ramos, M. (2021). Examining the use self-perceived by university teachers about ict resources: Measurement and comparative analysis in a one-way ANOVA design. *Contemporary Educational Technology, 13*(1), 1–13. https://doi.org/10.30935/cedtech/8707

Hastie, T., & Tibshirani, R. (2009). *K-fold cross-validation*. Available at: http://statweb.stanford.edu/~tibs/sta306bfiles/cvwrong.pdf. (Accessed 4 May 2020).

Hussain, S., & Khan, M. Q. (2023). Student-Performulator: Predicting students' academic performance at secondary and intermediate level using machine learning. *Annals of Data Science, 10*(3), 637–655. https://doi.org/10.1007/s40745-021-00341-0

Jimoyiannis, A., Koukis, N., & Tsiotakis, P. (2020). Shifting to emergency remote teaching due to the COVID-19 pandemic: An investigation of Greek teachers' beliefs and experiences. *Communications in Computer and Information Science, 1384 CCIS*, 320–329. https://doi.org/10.1007/978-3-030-73988-1_25

Kafedžić, E., Maleč, D., & Nikšić, E. (2018). Differences between male and female secondary school students in assessing their physical and health education teachers ' competences. *Sports Science, 11*(Suppl 1), 64–70.

König, J., Jäger-Biela, D. J., & Glutsch, N. (2020). Adapting to online teaching during COVID-19 school closure: Teacher education and teacher competence effects among early career teachers in Germany. *European Journal of Teacher Education, 43*(4), 608–622. https://doi.org/10.1080/02619768.2020.1809650

LaViale, T. (2023). *Deep dive on KNN: Understanding and implementing the K-nearest neighbors algorithm*. Available at: https://arize.com/blog-course/knn-algorithm-k-nearest-neighbor/. (Accessed 29 August 2023).

Leijon, M., Gudmundsson, P., Staaf, P., & Christersson, C. (2022). Challenge based learning in higher education– A systematic literature review. *Innovations in Education & Teaching International, 59*(5), 609–618. https://doi.org/10.1080/14703297.2021.1892503

Lubis, A. R., Lubis, M., & Al-Khowarizmi. (2020). Optimization of distance formula in K-Nearest Neighbor method. *Bulletin of Electrical Engineering and Informatics, 9*(1), 326–338. https://doi.org/10.11591/EEI.V9I1.1464

Maldonado, J., Riff, M. C., & Neveu, B. (2022). A review of recent approaches on wrapper feature selection for intrusion detection. *Expert Systems with Applications, 198*, Article 116822. https://doi.org/10.1016/j.eswa.2022.116822

Malik, T. (2011). *College success: First year seminar's effectiveness on freshmen academic and social integration, impact on academic achievement and retention at a southern institution*. In ProQuest LLC. Available at: https://0-search-proquest-com.biblioteca-ils.tec.mx/dissertations-theses/college-success-first-year-seminars-effectiveness/docview/964175051/se-2. (Accessed 17 April 2023).

Martinez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernandez-Orallo, J., Kull, M., Lachiche, N., Ramirez-Quintana, M. J., & Flach, P. (2021). CRISP-DM twenty years later: From data mining processes to data science trajectories. *IEEE Transactions on Knowledge and Data Engineering, 33*(8), 3048–3061. https://doi.org/10.1109/TKDE.2019.2962680

Mercader, C., & Gairín, J. (2020). University teachers' perception of barriers to the use of digital technologies: The importance of the academic discipline. *International Journal of Educational Technology in Higher Education, 17*(1), 4. https://doi.org/10.1186/s41239-020-0182-x

Mienye, I. D., & Sun, Y. (2022). A survey of ensemble learning: Concepts, algorithms, applications, and prospects. *IEEE Access, 10*, 99129–99149. https://doi.org/10.1109/ACCESS.2022.3207287

Mishra, A. (2018). *Metrics to evaluate your machine learning algorithm | towards data science*. https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234.

Mishra, P., & Koehler, M. J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. *Teachers College Record: The Voice of Scholarship in Education, 108*(6), 1017–1054. https://doi.org/10.1177/016146810610800610

Mukesh S, S. (2022). Outcome-based learning: An overview. *SSRN Electronic Journal*. https://doi.org/10.2139/SSRN.4026986

Müller, C., & Mildenberger, T. (2021). Facilitating flexible learning by replacing classroom time with an online learning environment: A systematic review of blended learning in higher education. *Educational Research Review, 34*, Article 100394. https://doi.org/10.1016/J.EDUREV.2021.100394

Muncie, T. (2020). *Using machine learning models to predict student retention: Building a state-wide early warning system*. Morehead State Theses and Dissertations. Available at: https://scholarworks.moreheadstate.edu/msu_theses_dissertations/868. (Accessed 9 March 2023).

Muntean, M., & Militaru, F. D. (2023). Metrics for evaluating classification algorithms. *Smart Innovation, Systems and Technologies, 321*, 307–317. https://doi.org/10.1007/978-981-19-6755-9_24/COVER

Nayak, P., Vaheed, S., Gupta, S., & Mohan, N. (2023). Predicting students' academic performance by mining the educational data through machine learning-based classification model. *Education and Information Technologies, 2023*, 1–27. https://doi.org/10.1007/S10639-023-11706-8

Ndukwe, I. G., & Daniel, B. K. (2020). Teaching analytics, value and tools for teacher data literacy: A systematic and tripartite approach. *International Journal of Educational Technology in Higher Education, 17*(1), 22. https://doi.org/10.1186/s41239-020-00201-6. Springer.

Ngo, G., Beard, R., & Chandra, R. (2022). Evolutionary bagging for ensemble learning. *Neurocomputing, 510*, 1–14. https://doi.org/10.1016/j.neucom.2022.08.055

OECD. (2019). *Indicator B5. How many students complete tertiary education?Education at a Glance 2019*. OECD. https://doi.org/10.1787/F8D7880D-EN. . (Accessed 9 March 2023)

OECD. (2022a). *How many students complete tertiary education? | education at a glance 2022 : OECD indicators | OECD iLibrary*. Available at: https://www.oecd-ilibrary.org/education/education-at-a-glance-2022_e3b05354-en. (Accessed 9 March 2023).

OECD. (2022b). *The assessment of higher education learning outcomes - OECD*. Available at: https://www.oecd.org/education/imhe/theassessmentofhighereducationlearningoutcomes.htm. (Accessed 7 November 2022).

Okoye, K. (2022). Using strategic intelligence and technology as building block for educational innovation: A conceptual framework towards the impact for outcome-based education. In *Elecom 2022 - proceedings of the 2022 4th IEEE international conference on emerging trends in electrical, electronic and communications engineering* https://doi.org/10.1109/ELECOM54934.2022.9965251

Okoye, K., Arrona-Palacios, A., Camacho-Zuñiga, C., Achem, J. A. G., Escamilla, J., & Hosseini, S. (2022). Towards teaching analytics: A contextual model for analysis of students' evaluation of teaching through text mining and machine learning classification. *Education and Information Technologies, 27*(3), 3891–3933. https://doi.org/10.1007/S10639-021-10751-5

Okoye, K., Arrona-Palacios, A., Camacho-Zuñiga, C., Hammout, N., Nakamura, E. L., Escamilla, J., & Hosseini, S. (2020). Impact of students evaluation of teaching: A text analysis of the teachers qualities by gender. *International Journal of Educational Technology in Higher Education, 17*(1), 49. https://doi.org/10.1186/s41239-020-00224-z

Okoye, K., Rodriguez-Tort, J. A., Escamilla, J., & Hosseini, S. (2021). Technology-mediated teaching and learning process: A conceptual study of educators' response amidst the covid-19 pandemic. *Education and Information Technologies, 26*(6), 7225–7257. https://doi.org/10.1007/s10639-021-10527-x

Olive, D. J. (2017). Multiple linear regression. In *Linear regression* (pp. 17–83). Springer International Publishing. https://doi.org/10.1007/978-3-319-55252-1_2.

Palacios, C. A., Reyes-Suárez, J. A., Bearzotti, L. A., Leiva, V., & Marchant, C. (2021). Knowledge discovery for higher education student retention based on data mining: Machine learning algorithms and case study in Chile. *Entropy, 23*(4), 485. https://doi.org/10.3390/E23040485

Peker, S., & Kart, Ö. (2023). Transactional data-based customer segmentation applying CRISP-DM methodology: A systematic review. *Journal of Data, Information and Management, 2023*, 1–21. https://doi.org/10.1007/S42488-023-00085-X

Ploutz, E. C. (2018). Machine learning applications in graduation prediction at the university of Nevada, las vegas. *UNLV Theses, Dissertations, Professional Papers, and Capstones*. https://doi.org/10.34917/13568668. . (Accessed 9 March 2023)

Priyambada, S. A., Usagawa, T., & Er, M. (2023). Two-layer ensemble prediction of students' performance using learning behavior and domain knowledge. *Computers and Education: Artificial Intelligence, 5*, Article 100149. https://doi.org/10.1016/j.caeai.2023.100149

Rstudio. (2023). *RStudio – statistics software*. Available at: https://rstudio.com/products/rstudio/. (Accessed 20 April 2020).

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science, 181*, 526–534. https://doi.org/10.1016/J.PROCS.2021.01.199

Shambour, M., & Abu-Hashem, M. (2022). Analysing lecturers' perceptions on traditional vs. distance learning: A conceptual study of emergency transferring to distance learning during COVID-19 pandemic. *Education and Information Technologies, 27*(1), 3225–3245. https://doi.org/10.1007/S10639-021-10719-5

Shearer, C. (2000). The CRISP-DM model: The new blueprint for data mining. *Journal of Data Warehousing, 5*(4), 13–22.

Shinde, P. P., & Shah, S. (2018). A review of machine learning and deep learning applications. In *Proceedings - 2018 4th international conference on computing, communication control and automation*. ICCUBEA. https://doi.org/10.1109/ICCUBEA.2018.8697857, 2018.

Sibindi, R., Mwangi, R. W., & Waititu, A. G. (2023). A boosting ensemble learning based hybrid light gradient boosting machine and extreme gradient boosting model for predicting house prices. *Engineering Reports, 5*(4). https://doi.org/10.1002/eng2.12599

Sun, F. R., Hu, H. Z., Wan, R. G., Fu, X., & Wu, S. J. (2019). A learning analytics approach to investigating pre-service teachers' change of concept of engagement in the flipped classroom. *Interactive Learning Environments, 0*(0), 1–17. https://doi.org/10.1080/10494820.2019.1660996

TEC. (2020a). *TEC, Flexible and Digital Model for academic continuity. NUVE Magazine*. Available at: https://www.revistanuve.com/modelo-flexible-y-digital-para-la-continuidad-academica/. (Accessed 3 August 2021).
TEC. (2020b). *Tec flexible digital plus model | tecnológico de Monterrey*. Available at: https://tec.mx/en/plus-digital-flexible-model. (Accessed 19 July 2020).
TEC. (2023a). *Challenge based learning*. Available at: https://tec.mx/en/challenge-based-learning. (Accessed 22 December 2023).
TEC. (2023b). *TEC21 Model - a new way of learning*. Available at: https://tec.mx/en/tec-model. (Accessed 22 December 2023).
Torres-Barreto, M. L., Castro-Castaño, G. P., & Melgarejo, M. A. (2020). A learning model proposal focused on challenge-based learning. *Advances in Engineering Education, 8*(2). Summer 2020.
Tzovla, E., Kedraka, K., & Kaltsidis, C. (2021). Investigating in-service elementary school teachers' satisfaction with participating in MOOC for teaching biological concepts. *Eurasia Journal of Mathematics, Science and Technology Education, 17*(3), Article em1946. https://doi.org/10.29333/EJMSTE/9729
Uliyan, D., Aljaloud, A. S., Alkhalil, A., Amer, H. S. Al, Mohamed, M. A. E. A., & Alogali, A. F. M. (2021). Deep learning model to predict students retention using BLSTM and CRF. *IEEE Access, 9*, 135550–135558. https://doi.org/10.1109/ACCESS.2021.3117117
UNESCO. (2015). *Competency based education. Learning portal - Planning education for improved learning outcome*. Available at: https://learningportal.iiep.unesco.org/en/library/competency-based-education. (Accessed 18 February 2020).
UNESCO. (2020). *Global education monitoring report 2020: Inclusion and education: All means all*. Paris: UNESCO. https://doi.org/10.54676/JJNK6989. . (Accessed 9 March 2023)
UNESCO. (2021). *National learning platforms and tools*. Available at: https://en.unesco.org/covid19/educationresponse/nationalresponses. (Accessed 18 August 2021).
UNESCO. (2022a). *Official list of SDG 4 indicators FFA education 2030 framework for action*. Available at: https://tcg.uis.unesco.org/wp-content/uploads/sites/4/2020/09/SDG4_indicator_list.pdf. (Accessed 7 July 2022).
UNESCO. (2022b). *Assessment for improved learning outcomes | UNESCO*. Available at: https://www.unesco.org/en/education/assessment. (Accessed 7 November 2022).
UNESCO. (2023). *Launch of the 2023 survey of formal education for SDG4 data | UNESCO UIS*. Available at: https://uis.unesco.org/en/news/launch-2023-survey-formal-education-sdg4-data. (Accessed 16 May 2023).
van der Aalst, W. M. P. (2016). Process mining: Data science in action. In *Process mining: Data science in action*. Springer https://doi.org/10.1007/978-3-662-49851-4.
Veenstra, C. P. (2009). A strategy for improving freshman college retention. *Journal for Quality and Participation, 31*(4), 19–23.
Viji, C., Beschi Raja, J., Ponmagal, R. S., Suganthi, S. T., Parthasarathi, P., & Pandiyan, S. (2020). Efficient fuzzy based K-nearest neighbour technique for web services classification. *Microprocessors and Microsystems, 76*, Article 103097. https://doi.org/10.1016/j.micpro.2020.103097
Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. In *4th int. Conference on practical application of knowledge discovery and data mining, manchester, UK* (pp. 29–40). http://cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf.
Wong, T.-T., & Yeh, P.-Y. (2019). Reliable accuracy estimates from k-fold cross validation. *IEEE Transactions on Knowledge and Data Engineering*. https://doi.org/10.1109/tkde.2019.2912815, 1–1.
Xiong, Z., Cui, Y., Liu, Z., Zhao, Y., Hu, M., & Hu, J. (2020). Evaluating explorative prediction power of machine learning algorithms for materials discovery using k-fold forward cross-validation. *Computational Materials Science, 171*, Article 109203. https://doi.org/10.1016/j.commatsci.2019.109203
Zhang, Y., Cao, G., Wang, B., & Li, X. (2019). A novel ensemble method for k-nearest neighbor. *Pattern Recognition, 85*, 13–25. https://doi.org/10.1016/j.patcog.2018.08.003
Zhao, Q., Wang, J. L., & Liu, S. H. (2021). A new type of remedial course for improving university students' learning satisfaction and achievement. *Innovations in Education & Teaching International, 59*(6), 711–723. https://doi.org/10.1080/14703297.2021.1948886