Advancements in Machine Learning Techniques for Educational Data Mining: An Overview of Perspectives and Trends

Rithesh Kannan[1*], Timothy Tzen Vun Yap[2], Hu Ng[3], Lai Kuan Wong[4], Fang Fang Chua[5], Vik Tor Goh[6], Yee Lien Lee[7], Hwee Ling Wong[8]

[1,3,4,5]*Faculty of Computing & Informatics (FCI), Multimedia University; E-mail: 1181301564@student.mmu.edu.my*

[2]*School of Mathematical & Computer Sciences, Heriot-Watt University, Putrajaya, Malaysia*

[6,7,8]*Faculty of Engineering (FOE), Multimedia University*

Abstracts: In the educational data mining (EDM) field, predicting student at-risk, student retention, dropout and performance have been attractive tasks among researchers. However, it is difficult to develop accurate models without first performing proper feature selection and class balancing. Therefore, the goal of this study is to review the current and future perspective and trends within the field of EDM for the past 10 years. The goal is to understand the state-of-the-art methods and techniques involving feature selection, class balancing and machine learning models. From the analysis, it is understood that there are plenty of research gaps yet to be explored.

Keywords: Educational Data Mining (EDM), Review, Feature Selection, Class Balancing, Machine Learning.

1. INTRODUCTION

Education is one of the most important aspects of a human life. This has been reflected by the increasing amount of concern for higher education institutes (HEI) on improving the quality of education and decreasedropout rates. This is because HEIs need to focus on efficiency and cut costs with the current trend of the economy that focuses on fierce competition. There are many measures of education quality, such as student performance, dropout rate, retention rate, graduating on time, employment status, and so on. In this paper, the focus is on reviewing the current research work that improves the quality of experience for students.

1.1. Educational Data Mining

Recently, there has been a field that is interested in creating methods to explore and find underlying patterns within educational datasets using machine learning (ML) and data mining (DM) techniques. The field is called educational data mining (EDM) and it is a discipline concerned with improving student performanceand their learning environment using ML techniques (Shafiq, Marjani, Habeeb, & Asirvatham, 2022).

In EDM, there has always been a great interest to predict ‘students at-risk’, which refers to those students whose performance is not good and have a high probability of dropping out or not graduating on time. This is because the knowledge can help educational institutes to provide an early warning to the affected students and decrease their dropout rate. However, attempting to measure and predict these ‘students at-risk’ is a challenging task without first managing the inherent class imbalance of the dataset and identifying the key features using features selection methods.

Therefore, there is a need to explore class balancing techniques and feature selection methods in EDM to explore the inconspicuous relationship in the data and reveal the patterns that will help predict students at-risk. There is also a need to understand the current state-of-the-art ML and DM techniques used in EDM research.

1.2. Class imbalance

Fernández et al. (2018) mentioned that class imbalance is a problem when general classification prediction techniques become biased towards the majority classes which causes there to be a higher misclassification rate for the minority class instances. There are three main perspectives of class balancing methods (CBM),data level, algorithm level and ensemble level.

- Data level: It is divided into under sampling (US) and over sampling (OS). Under sampling refers to when samples from the majority class are removed to make both classes equal. E.g.: Random US(RUS), Tomek-Link, One sided selection (OSS). Oversampling refers to when samples from the minority class are generated to make both classes equal. E.g.: Random OS (ROS), SMOTE, ADASYN.

- Algorithm level: It is divided into cost-sensitive learning and learning function modification. Cost-sensitive learning refers to methods that implement a cost function besides the normal loss function to an algorithm. E.g.: Cost-sensitive MLP, Cost-sensitive SVM. Learning function modification refers to adding a balancing function modification to the learning function of the algorithm itself. E.g.: Modified error BPNN.

- Ensemble level: It is divided into methods that combine with data level and methods that combine with algorithm level. Data level solutions include SMOTE-Boost and ROS-Bagging. Algorithm level solutions include MetaCost and Cost-sensitive XGBoost.

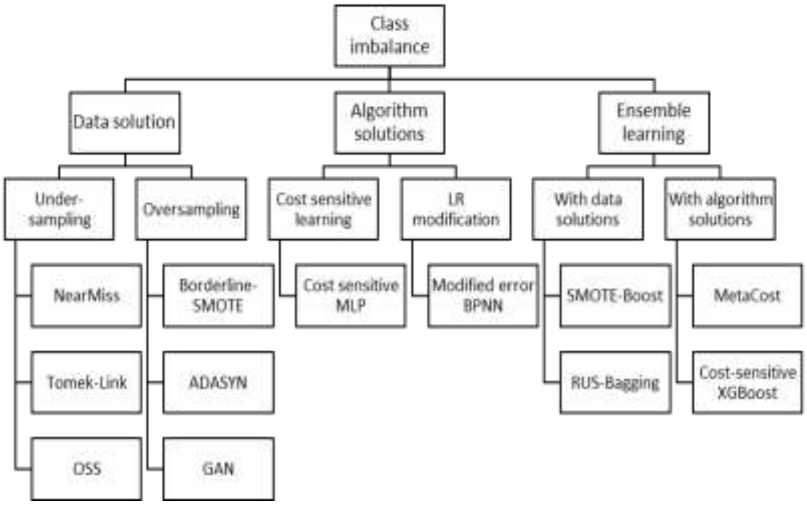

Figure 1: Class balancing methods.

1.3. Feature selection

Feature selection (FS) is a method used to select a minimum subset of features that are believed to be the most useful in the prediction task. There are three types of techniques for feature selection, filter, wrapper,and embedded methods (Ai, Zhang, Yu, & Shao, 2020).

- Filter method: Filter methods generally use statistical methods to evaluate the relationship between each input feature and the target feature. The scores are then used as the basis to select or filter the input features to be used in the model. They are most generally used in the preprocessing stage and are faster in computing than wrapper methods.

- Wrapper method: Wrapper methods are used to create multiple models, each with a different subset of the input features. The method then selects the features that lead to the optimal model with the best performance according to the performance metric. These methods are not concerned with the variable types and are generally more computationally expensive than filter methods but perform better in getting the best subset of input features.

- Embedded method: These methods combine the advantages of the previous two methods, the speed of filter methods and the performance of wrapper methods and embed it into the learning algorithmitself. These methods are also iterative which allows it to optimally find the most important featuresthat contribute the most in each iteration.

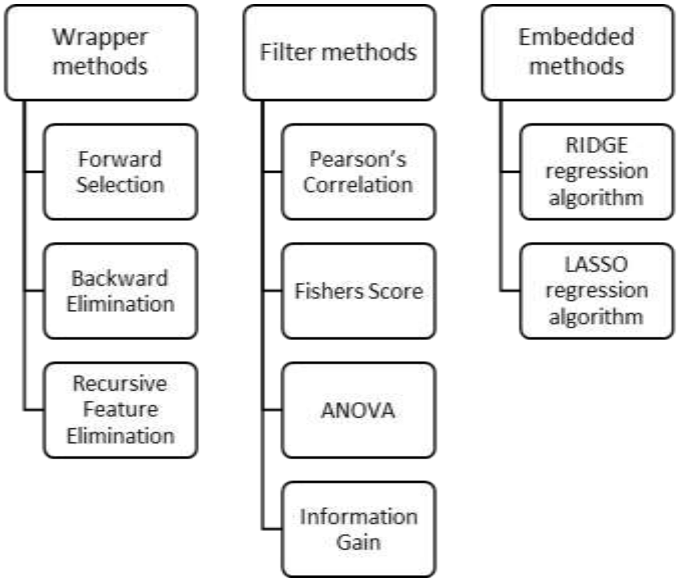

Figure 2: Feature selection methods.

1.4. Machine learning (ML) methods

According to Awad & Khanna (2015), machine learning is a subfield under Artificial Intelligence (AI) andis concerned with systematically applying algorithms that synthesize the underlying relationship between data and information. There are three main ways ML models are implemented: supervised, unsupervised and ensemble.

- Supervised method: It is a ML approach that consists of using labeled datasets to train or "supervise" algorithms into predicting outcomes or classifying data accurately. The model can measure its accuracy using the labeled input features and the target output feature and learn over time to improveits performance.

- Unsupervised method: It used ML algorithms to cluster and analyze unlabeled data sets. The goal is to discover the underlying, hidden relationships in data without the need for labelling or human intervention (hence, they are "unsupervised")

- Ensemble method: It is a combination of several ML models into one framework. Each of the models comprising the framework are known as weak learners, and the output of one serve as the input to the next model in sequence. The idea is to combine several weak learners into becoming strong learners.

The purpose of this research is to provide a summary of recent research in the EDM field for the past 10 years. The organization of the rest of the paper is shown below. Section 2 covers the related work where review papers covered by other researchers in the field of EDM are covered. Section 3 explains the methodology by which this review has been conducted which includes the data search strategy, the framework used to create the research questions, and finally the research questions and objectives for this review. Section 4 will include the results of the analysis done on the surveyed articles and the key findings are presented. Section 5 will present the discussion centered on the reviewed articles and showcase the overall trend and key perspectives in the field of EDM through the research gaps. Finally, in Section 6, a summary of the current review paper is provided, along with suggestions for future work.

2. Related Works

This section explains other review papers in the field of EDM by summarizing and comparing their important findings.

There have been plenty of reviews in the field of EDM that cover different aspects. 11 different types of review papers have been covered over the past 5 years by the authors. There are many different types of review papers such as narrative review, systematic review, theoretical review, overview review, and so on (Grant & Booth, 2009).

Narrative reviews, also known as traditional reviews, are used to provide an overview of a particular topic. They examine current or recent literature and employ a narrative approach to reporting the review findings. They typically do not list their inclusion/exclusion criteria, which may lead to them containing bias. Systematic reviews, on the other hand, will typically list their inclusion/exclusion criteria and seeks to systematically search for, appraise and synthesize research articles by following a guideline/protocol. Theoretical review, or qualitative reviews, are used to integrate or compare the findings from qualitative studies. They look for ‘themes’ or ‘ideas’ that lie across in individual research articles. Finally, an overview review is used to perform a summary of the related research articles by surveying the literature and describing their characteristics and research trends.

To the best of the author’s knowledge, other selected reviews have limited coverage on the topic of implementing feature selection and class balancing techniques together with machine learning models. Existing reviews are concerned with understanding the factors the determine student performance as well as the machine learning models and algorithms that help predict them. While the existing reviews provide interesting findings and some future research directions, they also have different limited scope.

Some of the ways in which the scope is different from the current review are:

- Some review papers do not provide details about data collection methods or factors used (Nik Nurul Hafzan, Safaai, Asiah, Mohd Saberi, & Siti Syuhaida, 2019; Ranjeeth, Latchoumi, & Paul, 2020).
- The review process of some papers reveals a lack in more holistic approaches to research the studentexperience (Tight, 2019).
- Some review papers covered only a limited number of papers (Ranjeeth et al., 2020) or they were limited by their research sources (Du, Yang, Hung, & Shelton, 2020), or by their data source, and the years they selected to review (Alyahyan & Düştegör, 2020).

Table 1: Summary of selected review papers, sorted by year.

No.	Ref & Year	Review Type	Studies reviewed	Finding(s)
1	Nik Nurul Hafzan et al., 2019	Narrative Review	50	Compares Dynamic Bayesian network versus other ML models, to predict at-risk students and decrease drop-out rates.
2	Tight, 2019	Theoretical Review	4344 from 1960 – 2018	Aims to determine the origin and meaning behind the terms, 'student retention' and 'student engagement'.
3	Ranjeeth et al., 2020	Overview review	30 from 2015-2020	Showcases the progress of learning analytics in the past five years, in terms of prediction algorithms, datasets used, and prediction factors.
4	Rastrollo-Guerrero, Gómez-Pulido, & Durán-omínguez, 2020	Narrative Review	64 from 2013-2019	Reviews the state-of-the-art techniques in predicting students' performance as well as the objectives that researchers must reach in the field.
5	Du et al., 2020	Systematic Review	1219 from 2007-2019	Reviews the current research trends by finding out the primary research topics and findings, as well as any open issues and future trends in EDM research.
6	Alyahyan & Düştegör, 2020	Narrative Review	2015-2020	Provides a set of guidelines step-by-step for educators willing to apply DM techniques for student success prediction by comprehensively covering possible decisions and parameters along with arguments.
7	Rodrigues, Dos Santos, Costa, & Moreira, 2022	Systematic Review	24 from 2012-2021	Reviews the current state-of-the-art models used to predict student performance in high school and elementary school.
8	Shafiq et al., 2022	Systematic Review	100 from 2017-2021	Shows what factors help to determine the high risk-students as well as what learning analytics and machine learning approaches are used to improve the performance for students and teaching practices.
9	Saluja & Rai, 2022	Systematic Review	40	Highlights the techniques that are being used to do prediction of students' academic performance and the advantages and disadvantages of each method found in various research works.
10	Issah, Appiah, Appiahene, & Inusah, 2023	Systematic Review	84 from 2016-2022	Reviews current research studies to find the most frequently applied characteristics, ML methods and algorithms that are used to do prediction of student performance.
11	Abdul Bujang et al., 2023	Systematic Review	43 from 2015-2021	Reviews latest approaches that handle higher education imbalanced classification, which includes the current practices of data features, techniques, and a comparative analysis of the proposed algorithms, and focuses on predicting student performance.

3. METHODOLOGY

3.1. Data Search

The papers were collected from various database sources shown in Figure 4. The 'Others' sections included sources like MECS (International Journal of Modern Education and Computer Science), SAI (The Science and Information) Organization, Informatica, BMC (Bio Med Central) medical education IJIET

(International Journal of Information and Education Technology), IJASCA (International Journal of Advances in Soft Computing and its Applications), JTEC (Journal of Telecommunication, Electronic and Computer Engineering), MATEC Web of Conferences, Taylor and Francis. A total of 66 papers were selected, including other review papers and papers with focus on class imbalance, feature selection and machine learning.

Figure 3 shows the number of published articles by year for the selected articles. From the below figurethe trend of papers being published in the field of EDM is increasing in general. There may be less papers in 2023 as it is the current year. Figure 4 shows the data sources from which the articles were collected from. It is shown that the most popular publication source among the reviewed papers is IEEE Access (30%), followed by Science Direct (28%).

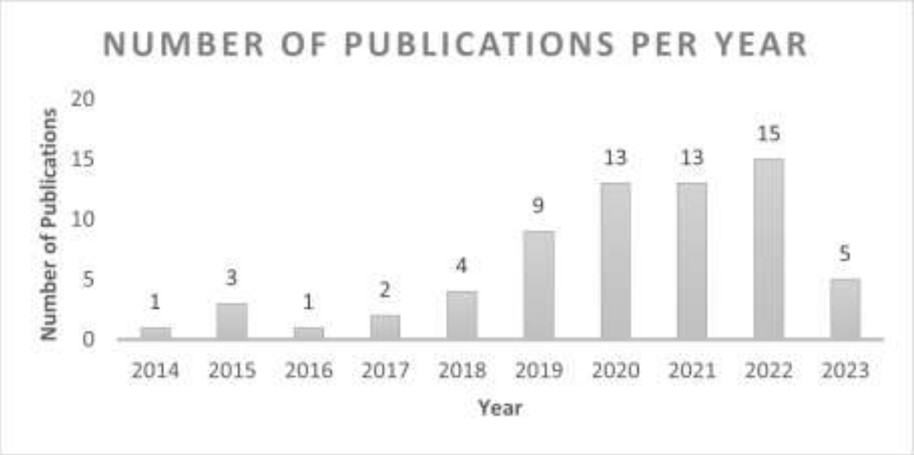

Figure 3: Number of publications reviewed by year.

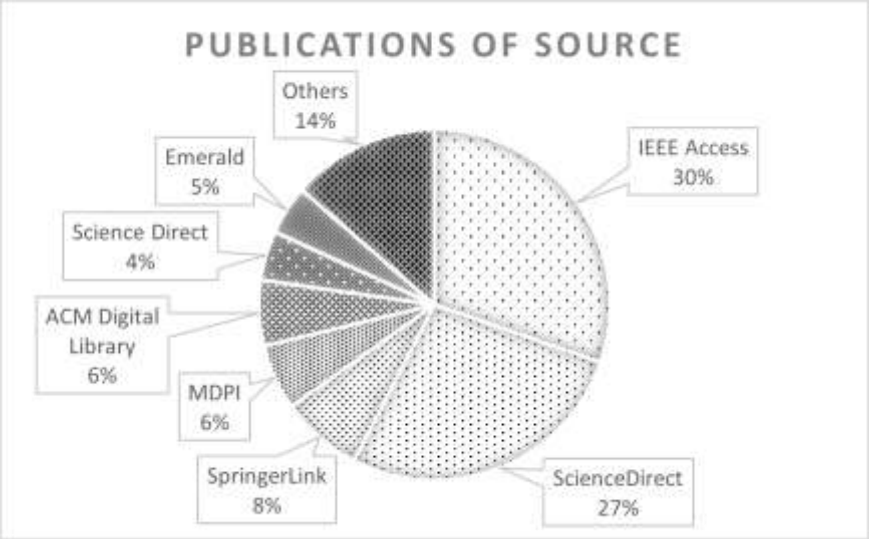

Figure 4: Publications by source.

3.2. Framework

For this review paper, the SPIDER framework has been used to generate research questions and formulate the search strategy (Cooke, Smith, & Booth, 2012). The SPIDER framework is a tool that is usedas a systematic strategy in review papers to search the articles and narrow down the research questions that this review paper will answer. The step-by-step process is listed below:

I. S-Sample: This refers to the group of the population that are being studied. In our case, the students'experiences and their performances are investigated.

II. P of I-Phenomenon of Interest: This refers to the topic of research or intervention. In this case, the focus is on class balancing, feature selection and machine learning techniques in EDM research forthe past 10 years.

III. D-Design: This refers to the techniques used to gather data. In this paper, the state-of the arttechniques, methods, and outcomes in EDM research are gathered and observed.

IV. E-Evaluation: It refers to the outcome of the study, which in this case examines the overall trendand perspectives of current research in EDM.

V. R-Research Type: It refers to what kind of qualitative study has been conducted. For this paper, aqualitative review has been conducted.

For this review paper, four research questions have been formulated using the SPIDER framework to narrow down the scope and to help search and identify the relevant articles. The four research questions are listed below:

I. What are the factors that can help predict students at-risk? To discover the common factors from the reviewed papers that help predict students at-risk so that the underlying relationship between the factors is understood. From this, the current focus of researchers can also be seen and any factors that were overlooked can be seen.

II. What are the current class balancing approaches that are used to improve the prediction of studentsat-risk? To identify the common class balancing approaches to see where they are utilized and whatis the most popular approach being used.

III. What are the current feature selection techniques that are used to improve the prediction of studentsat-risk? To understand the commonly used feature selection techniques to see which is the most popular technique as well as what are the factors/features generally being ranked as important.

IV. What are the current machine learning models that are used to predict students at-risk with high accuracy? To review and compare the common machine learning models through three perspectives, supervised, unsupervised and ensemble approach and ascertain the future trends in EDM field.

This review paper has undertaken an extensive search process based on the research questions and the SPIDER framework. The search terms used to extract the relevant papers are given below based on three considerations: A) The technique of the paper or what models they have used, B) The purpose of thepaper, and C) What was their focus of the paper in relation to the research questions, that is whether they used class balancing or feature selection or supervised and unsupervised ML techniques. In Table 2 the eligibility criteria for including and excluding papers can be seen.

Search Terms:

A) Technique: (("Machine learning" OR "Educational data mining (EDM)" OR "Predictive model*"OR "Learning analytic*")

B) Purpose: (AND "Academic performance" OR "Student dropout prediction" OR "Identifying at-riskstudent*" OR "Student retention" OR "Student performance")

C) Method: (AND " Class *balance*" OR "Feature selection" OR "Unsupervised" OR "Supervised")

Table 2: Inclusion and exclusion criteria for papers.

Criteria	Inclusion Criteria	Exclusion Criteria
Years	2013-2023	Any articles before 2008
Language	English	Languages besides English
Publication Status	Fully Published	Articles not fully published

4. REVIEW ANALYSIS

In this review on the trend of EDM, three perspectives are considered, which are class balancing, feature selection and machine learning. The reason is because the goal for all the perspectives is the same, which is to improve the model performance and ensure that the results are effective in dealing with the needs of students and educational institutions. For each section, a table with a summary of all the important findingsis listed.

4.1. Class balancing methods

Jahin et al. (2021) used a custom technique based on oversampling called semi-supervised oversampling technique (SSOT) to solve class imbalance in their datasets and predict student performance. Pratama, Pristyanto, & Prasetyaningrum (2021) compared multiple CBMs like SMOTE, Borderline SMOTE, and SMOTE-Tomek and found the best method to be SMOTE-Tomek. Palacios, Reyes-Suárez, Bearzotti, Leiva, & Marchant (2021) compared multiple models before and after class balancing using SMOTE and found RF model to be the best performing, with accuracy exceeding 80%. Alwarthan, Aslam, & Khan (2022) compared under sampling, oversampling and hybrid methods through SMOTE, Tomek-links and SMOTE- Tomek Link respectively, and found RF to once again be the best performing model. Masood & Begum (2022) compared five different OS techniques with two ML models and found the two best performing models to be SVM with RUS and SMOTE with AUC-Score of 0.69 and 0.70, respectively.

Sha, Rakovic, Das, Gasevic, & Chen (2022) compared fairness and accuracy of the models using three class balancing techniques (CBTs) each from under sampling, oversampling, and hybrid methods. They found the three best methods to be SMOTE, SMOTE-Tomek Link, and Tomek-Link that improved fairnessof the model in all scenarios. Verma, Yadav, & Kholiya (2022) included an algorithm balancing method in comparison alongside data balancing methods. They used SVM-SMOTE, and compared it with regular SMOTE, Borderline SMOTE and ADASYN. Dileep, Bansal, & Cunningham (2022) compared multiple models, including ensemble models like CatBoost before and after applying SMOTE. Cabezuelo, González, Campo, Barbero, & Mduma (2023) also compared models with five data balancing methods which were: SMOTE, SMOTE ENN, RUS, ROS, and SMOTE TOMEK. Out of them, SMOTE ENN performed the best.Alija, Beqiri, Gaafar, & Hamoud (2023) predicted student performance by comparing multiple models with SMOTE as the balancing method. Once again, RF performed the best.

In Table 3, the imbalance ratio (IR) for each paper is also listed. The IR is a measure used to describe the level of imbalance within a dataset (Zhu, Guo, & Xue, 2020). It can be defined as:

$$IR = \frac{S_{mj}}{S_{mn}} \tag{1}$$

where Smj refers to the sample size of the majority class and Smn refers to the sample size of the minority class. In case there are multiple classes, the Smj will be the size of the largest class within the sample and the Smn will be the size of the smallest class within the sample. In summary, when IR < 1, the dataset is balanced. When IR > 1, the dataset is imbalanced. The larger the IR, the larger the severity of imbalance within the dataset.

From the articles reviewed, many researchers choose to focus on data level solutions to class balance, specifically oversampling as all contain at least one oversampling technique. Four papers contain under sampling technique and only one paper uses algorithm level class balancing. Five papers have used ensemble level class balancing by combining oversampling and under sampling techniques together.

Table 3a: Summary of selected class balancing focused articles, sorted by year.

No	Ref & Year	Imbalance Ratio (IR)	Class balancing method (CBM)	Finding(s)
1	Jahin et al., 2021	IR=85%/15%=5.6	Semi-supervised OS Technique (SSOT).	Proposed a semi-supervised OS method to balance dataset and predict student performance in given course.
2	Pratama et al., 2021	IR=211/126=1.7	SMOTE, Borderline SMOTE, SMOTE-Tomek Link.	Shows the effect of the imbalanced data problem and compares resampling methods to find optimal one to be implemented into the ML process.
3	Palacios et al., 2021	IR=33:1=33	SMOTE	Formulates EDM models based on ML techniquesto extract appropriate information from educational data in HEI and utilizes the information in the knowledge discovery in databases (KDD) process.
4	Alwarthan et al., 2022	Not mentioned	SMOTE, Tomek Link, SMOTE-Tomek Link.	Predicted students at-risk at an early stage by applying several EDM models to build three classification models.
5	Masood & Begum, 2022	IR=18200/899=20.2	Random OS, Random US, SMOTE, SMOTE-ENN, SMOTE-Tomek Link.	The study presented the findings of comparing various resampling techniques used to address the problem of imbalanced data at the data preprocessing stage.
6	Sha et al., 2022	IR=2233/1470=1.5	NearMiss, Edited-NN, Condensed-NN, SMOTE, ADASYN, OSS, Ensemble CBMs, etc.	Investigated the performance of several CBMs on prediction fairness and applied hardness and distribution bias metrics to measure data characteristics that might have algorithmic bias.
7	Verma et al., 2022	IR=363/187=1.9	SMOTE, Borderline SMOTE, SVM-SMOTE, and ADASYN.	Predicted low performing students by identifying influential features and comparing five ML models with and without various CBMs.
8	Dileep et al., 2022	Not mentioned	SMOTE	Proposed a method for automatically detecting students at-risk by using online activity data in conjunction with student information system (SIS) data from a math course in a specific university.

Table 3b: Summary of selected class balancing focused articles, sorted by year.

9	Cabezuelo et al., 2023	IR=60359/981=61.5	RUS, ROS, SMOTE, SMOTE ENN, and SMOTE-Tomek Link.	Explored the use of various CBMs to improve the accuracy of ML models to predict student dropout.
10	Alija et al., 2023	IR =74/6=12.3	SMOTE	Student performance was predicted using supervised ML models on an imbalanced dataset. In addition, wrapper feature selection method was applied to select the most relevant features for the task of prediction

4.2. Feature selection methods

Punlumjeak, Rachburee, & Arunrerk (2017) was focused on predicting the performance of students using big data with the Microsoft Azure platform. Chung & Lee (2019) on the other hand, investigated predicting student dropout using the scaled difference in accuracies with and without the specific feature as their feature selection method. Olaya, Vásquez, Maldonado, Miranda, & Verbeke (2020) continued the trend of predicting student dropout, but they focused on high school students instead of university students. They used a method called the modified covariate approach (MCA) which showed the important features were socio-economic in nature such as gross family income.

Other researchers often experimented with wrapper and ensemble feature selection methods (Ai et al., 2020). They focused on producing a prediction framework that consisted of an ensemble feature selection method that itself consisted of various feature selection techniques linked in an elimination voting framework. Bello et al. (2020) used a Random Forest algorithm itself as a feature selection algorithm and used mean decrease accuracy (MDA) as the measure to eliminate features. Maldonado, Miranda, Olaya, Vásquez, & Verbeke (2021) combined filter based and wrapper-based methods into an ensemble framework. They used Fisher's score as the filter-based method and a backwards wrapper with logistic regression (LR) as the base classifier to eliminate variables.

Assistant, Nidhi, Majithia, & Sharma (2021) compared three filter-based feature selection methods called Correlation Attribute Evaluator (CAE), Information Gain Attribute Evaluator (IGAE), and Gain Ratio Attribute Evaluator (GRAE), to select the important predictors which were student grades, social, demographic, and school-related. Alraddadi, Alseady, & Almotiri (2021) applied their feature selection method to the two best performing ML models, logistic regression (LR) and linear discriminant analysis (LDA) to predict student performance as well. They used Binary Teaching-Learning Based Optimization (BTLBO) in a wrapper method with a V-shaped transfer function (TF). Arif, Jahan, Mau, & Tummarzia (2021) compared multiple wrapper methods in Weka and their best method was with a Random Forest algorithm. Nuanmeesri, Poomhiran, Chopvitayakun, & Kadmateekarun (2022) on the other hand used only filter based methods like Chi-Square, Gain Ratio, and Correlation-based Feature Selection (CFS) models to select the important features to predict student dropout. The features obtained from CFS model were student performance features like cumulative grade point average (CGPA) and socio-economic features like Loan status.

From the articles reviewed, there are four articles that use filter methods, two that use wrapper methods and four that use embedded methods that combine filter and wrapper. Filter methods seem to be more popular than wrapper methods from the selected articles.

Table 4: Summary of selected feature selection focused articles, sorted by year.

No	Ref	Important Features	Feature selection method	Finding(s)
1	Punlumjeaket al., 2017	Using mutual information: Student subjects' features	Filter method: Chi-square, Person correlation, Mutual information.	Presented several FS methods to find the best accuracy of the models to predict student performance.
2	Chung & Lee, 2019	Student behavior features	Wrapper method: Scaled difference in the accuracies.	Predicted high school student dropout using RF model with relevant data selected using FS method.
3	Olaya et al., 2020	Socio-economic features	Filter method: MCA	Proposed a new framework to prevent student attrition by utilizing uplift modeling.
4	Ai et al., 2020	Not mentioned	Embedded method: chi-square check, mutual information, T-test, MaxDiff, ReliefF	Proposed an integrated framework to predict student dropout which includes feature generation module and ensemble FS module.
5	Bello et al., 2020	Student performance features	Embedded method: RF algorithm and removing features with lowest MDA.	Identified the features with highest predictive value for student dropout using ML methods.
6	Maldonado et al., 2021	Not mentioned.	Embedded method: Fisher score, backward wrapper approach using LR.	Designed a new performance measure to evaluate predictive models for student dropout. The measure also quantifies the net savings achieved through a retention campaign.
7	Assistant et al., 2021	Students' grades, social, demographic, and school-related features	Filter method: CAE, IGAE and GRAE.	Compared the prediction result of base classification models with the classification models used with some FS methods.
8	Alraddadi et al., 2021	Not mentioned.	Embedded method: BTLBO ina wrapper method with V- shaped TF.	Introduced a hybrid framework to predict student performance using preprocessing techniques like FS along with well-performing ML models.
9	Arif et al., 2021	Demographic, socio-economic conditions, personal characteristics, marital status etc.	Wrapper method: WrapperSubsetEval in Weka took	Proposed a student performance prediction system that includes FS method and classification model.
10	Nuanmeesriet al., 2022	Using CFS model: Student performance factors and Socio-economic factors.	Filter method: Chi-Square, Gain Ratio, and Correlation- based Feature Selection (CFS) models.	Investigated the factors that affect student dropout and improved model performance by combining FS with MLP model.

4.3. Machine Learning methods

Litalien & Guay (2015) designed a model to understand the reasons for why PhD student's dropout. They used self-determination theory (SDT) and found perceived competence to be the best predictor for dropout intentions. For their future work, they have stated that they wish to extend their range to collect the data as their current range has some limitations.

Almasri, Alkhawaldeh, & Çelebi (2020) proposed a unified approach to build a new cluster-based (CB) classifier model, which groups together historical records of students into a set of homogenous clusters. Classifier models are then built for each cluster whose output along with the centroids of each cluster are fed into the final unified classifier model. A high level of accuracy of 96.25% was achieved. In future, theyhave stated that they will investigate hierarchical clustering techniques, as well as taking student behaviour such as anxiety and fatigue into consideration.

Chui, Fung, Lytras, & Lam (2020) proposed a reduced training vector-based support vector machine (RTV-SVM) model that modifies the original SVM model by removing the redundant training vectors with the aim of lowering the training time without lowering the accuracy. The model was able to predict at-risk students with 92.2 to 93.8% accuracy. For the future direction, they have stated they wished to test their model performance in other learning analytic applications.

Crivei, Czibula, Ciubotariu, & Dindelegan (2020) investigated two methods, principal component analysis (PCA) and relational association rule mining, to analyse student academic performance data and uncover meaningful patterns. They discovered that the potential for unsupervised models in detecting hidden patterns within student academic performance is quite high but noted that performance was influenced by anomalies and the small data sample size. For their future work, they planned to investigate methods to detect anomalies and outliers to reduce their impact on the performance.

From the selected articles, it is seen that many researchers choose to use ensemble models like Random Forest. There are ten researchers who chose to use only supervised models and two researchers who chose to use only unsupervised models. Three researchers chose to use other models due to their scope. Ensembleand supervised models are the popular models to use in EDM field.

Table 5a: Summary of selected machine learning focused articles, sorted by year.

No	Ref & Year	Finding(s)	Models
1	Rogers, Colvin, & Chiera, 2014	Compared the index method with linear multiple regression approach to identify students at risk and found that index method is comparable in performance.	Supervised: Index method, Standard linear multiple regression model.
2	Litalien & Guay, 2015	Provided a better understanding of factors contributing to PhD completion using SDT. The best predictor of dropout intentions was found to be perceived competence.	Others: Self-determination theory (SDT).
3	Lakkaraju et al., 2015	Developed a ML framework to identify high school students at-risk, by applying several classification algorithms (out of which RF performs the best) and evaluating metrics important to school administrators.	Ensemble: Adaboost (AB), Logistic Regression (LR), Random Forests (RF), Decision Trees (DT), and Support Vector Machines (SVM).

Table 5b: Summary of selected machine learning focused articles, sorted by year.

4	Jia & Maloney,2015	Influence of specific factors that may lead to non-completions and non-retentions were estimated and the effectiveness of predictive risk tool were tested to identify vulnerable students in an early manner.	Supervised: Maximum likelihood probit analysis.
5	Marbouti, Diefes-Dux, & Madhavan, 2016	Predictive models were compared to identify students at-risk in a course that used standards- based grading. The Ensemble and NBC model had the best results.	Ensemble: LR, SVM, DT, Multi-Layer Perceptron (MLP), Naïve Bayesian Classifier (NBC), KNN, Ensemble.
6	Azcona & Smeaton, 2017	Presented a system that uses ML methods by combining effort predictors and engagement to classify students in a course at an earlier stage.	Ensemble: LR, SVMs with linear and Gaussian kernels, DT, KNN classifier and RF.
7	Chen, Johri, & Rangwala, 2018	Survival analysis approaches were compared to several ML approaches to identify students at-risk of dropping out.	Ensemble: Cox's Proportional Hazard model, Aalen's Additive model, LR DT, RF, NBC and AdaBoost.
8	Miguéis, Freitas, Garcia, & Silva, 2018	Proposed a two-stage model that was supported byDM techniques to predict overall academic performance with accuracy above 95%	Ensemble: NBC, SVM, DT, RF, Bagged trees, Adaptive boosting trees (AdaBoost).
9	Shelton, Yang, Hung, & Du, 2018	An analytic approach was proposed that merges two predictive models (model of successful students and students at-risk) for the enhancement of prediction accuracy.	Ensemble: Neural network (NN), LR, RF, SVM and DT.
10	Burgos et al., 2018	Knowledge discovery methods were used to analyze historical data of student course grade to predict student dropout in a course.	Supervised: LOGIT Act, SEDM, FFNN, PESFAM, SVM.
11	Offiah et al., 2019	The retention levels of practical skills taught and assessed by SBE were evaluated and the extent of retraining needed to restore deteriorated performance was investigated.	Others: Observational prospective cohortstudy.
12	Oliveira, Barwaldt, Pias, &Espindola, 2019	The development and validation of predictive models that will be integrated into an early identification system were introduced. These models will identify students at-risk in courses.	Ensemble: KNN, NBC, C-Support Vector Classification (SVC), LR, RF, Gradient Boosting and Extremely Randomized Trees (Extra Trees), AdaBoost.
13	Azcona, Hsiao, & Smeaton, 2019	Implements a framework called PredictCS that detects students at-risk for three computer programming courses and adaptively and automatically sends them feedback.	Ensemble: KNN, Linear SVM, DT, RF.
14	Hill, Fulcher, Sie,& De Laat, 2019	Presented a work-in- progress in which an ensemble-based ML approach predicts the commencing students, while a simple logistic regression method predicts the ongoing students.	Ensemble: RF and LR.
15	Adekitan & Salau, 2019	Predictive analysis was carried out to determine final CGPA of engineering students using programof study, year of entry and GPA as inputs into KNIME based data mining model.	Supervised: PNN, RF, DT, NBC, Tree Ensemble Predictor, LR, Linear regressionmodel, Pure quadratic regression model.

Table 5c: Summary of selected machine learning focused articles, sorted by year.

16	Inyang, Eyoh, Robinson, & Udo, 2019	Adopted HCA to analyze students' performanceto discover the optimal number of clusters of failed courses and ARM for the extraction of interesting course-status association.	Unsupervised: Hierarchical Cluster Analysis (HCA), Association Rule Mining (ARM).
17	Moreno-Marcos et al., 2020	SRL strategies were analyzed on whether they can enhance existing predictive models for dropout, with or without common self-reported variables and variables derived from click-stream data.	Ensemble: Generalized Linear Model (GLM), SVM, RF, and DT.
18	Coussement, Phan, De Caigny, Benoit, & Raes, 2020	Improved student dropout predictions by benchmarking the LLM against eight other algorithms using a real-life dataset of a global subscription-based online learning provider.	Ensemble: Logit leaf model (LLM), LR, BOOST, SVM, RF, Logistic model tree (LMT), BAG, ANN, DT.
19	Chui et al., 2020	Proposed a RTV-SVM model designed to remove redundant training vectors to reduce thetraining time and preserve the support vectors, without sacrificing the classification accuracy.	Supervised: Reduced training vector-based support vector machine (RTV-SVM).
20	Kumar, Krishna, Neelakanteswara, & Basha, 2020	Applied clustering and classification methods toa dataset of students to evaluate student performance.	Ensemble: KNN clustering, Hierarchical clustering, DT and NBC.
21	Almasri et al., 2020	Proposed a unified framework to build a novel supervised cluster-based (CB) classifier to predict student performance.	Ensemble: KNN clustering, Functional MLP Probabilistic NBC, DT (J48) and Ensemble Meta-based Tree (EMT) model.
22	Crivei et al., 2020	Investigated if two unsupervised models, PCA and RAR mining, can identify patterns for predicting student final examination grade.	Unsupervised: Principal component analysis (PCA) and Relational association rule (RAR) mining (DRAR).
23	Alsharari & Alshurideh, 2021	Introduced a novel retention model designed for the academic setting that is based on the interplay between emotional intelligence, creativity, and learner autonomy.	Supervised: Smart Partial Least Square (SPLS).
24	Premalatha & Sujatha, 2021	Proposed to develop an ensemble of learning models to predict the employment status of graduates.	Ensemble: NBC, RF, Decision Stump (DS), DT (J48), MLP, Bagging, Ensemble.
25	Alcaraz, Martinez-Rodrigo, Zangroniz, & Rieta, 2021	Designed a tailored early warning system (EWS)for a conventional course in power electronic circuits, using ensemble classifier to predict at-risk students.	Ensemble: LR, NBC, DT, SVM, MLP, KNN, RF, AdaBoost, majority voting ensemble (MVE) approach.
26	Kuzilek, Zdrahal, & Fuglik, 2021	Predicted student academic outcomes by utilizing multiple predictive models that incorporate student exam behavior represented by the order of their exams.	Ensemble: Classification and Regression tree (CART), Non-linear Support Vector Machines with Radial Basis Function kernel (SVM-RBF). KNN, and RF

Table 5d: Summary of selected machine learning focused articles, sorted by year.

27	Mubarak, Cao, & Hezam, 2021	Proposed a hyper-model called CONV-LSTM, to automatically extract features from raw data in MOOCs. The model aims to perform student dropout prediction.	Supervised: Convolutional Neural Networks and Long Short-Term Memory (CONV- LSTM) with custom loss function, CONV- LSTM without custom loss function, Deep Neural Network (DNN), SVM, LR.
28	Paideya & Bengesai, 2021	The factors influencing persistence amongst students enrolled in a Chemistry major at a South African university were examined using enrolment data.	Supervised: CART DT.
29	Niyogisubizo, Liao, Nziyumva, Murwanashyaka, & Nshimyumukiza, 2022	Proposed a novel stacking ensemble based on a hybrid of ML models to predict student's dropout in university classes.	Ensemble: RF, XGBoost, Gradient boosting, FFNN.
30	Guzmán-Castillo et al., 2022	The implementation results of a predictive information system (IS) that prevents university student dropout were showcased	Ensemble: AdaBoost, Bayesian GLM, DT, Logit Boost, RF, and Stochastic Gradient Boosting.
31	Alboaneen et al., 2022	A web-based system was developed to predict student performance and identify students at-risk using several ML models.	Ensemble: SVM, RF, KNN, ANN, and LR.
32	Mariano, De Magalhães Lelis Ferreira, Santos, Castilho, & Bastos, 2022	Performed classification via decision trees to predict student evasion using data from Engineering course students in Brazil.	Supervised: DT (C4.5 algorithm).
33	Borrella, Caballero-Caballero, & Ponce-Cueto, 2022	A theoretical framework was proposed to provide guidance on the design of interventions aimed to reduce dropout rates in MOOCs.	Supervised: LR and RF.
34	Czibula, Ciubotariu, Maier, & Lisei, 2022	Introduced a generic ML-based framework called IntelliDaM, that aimsto improve the performance of data mining tasks and enhance the decision-making processes.	Ensemble: KNN clustering, Tweedie regressor (Generalized Linear Model (GLM), Stochastic Gradient De- scent (SGD), Polynomial regressor.
35	Jiang, Liu, Zhang, & Wang, 2023	A hybrid profit-driven customer churn prediction model was proposed that considers both return and cost.	Others: Modified multi-objective atomic orbital search.

5. DISCUSSIONS

In this section, there will be a discussion on the current research gaps that were listed in Section 4. Themost common features used to predict student performance are also listed.

The features that contribute to student performance are:

- Academic factors: It includes features like: Student subjects/course grades (Punlumjeak et al.,2017)

- Time related factors: This includes variables such as: Student lateness and absence (Chung & Lee,2019)

- Demographic/Socio-demographic factors: Features such as: Parents information such as employment (Olaya et al., 2020).

From the reviewed papers, the overall trend as well as future research directions can be observed. Someof those future directions are:

- Increase the sample size and the descriptive features: The main issue that many researchers have stated is the limited size of the data sample or the features (Alraddadi et al., 2021; Chung & Lee, 2019; Dileep et al., 2022; Jia & Maloney, 2015). The dataset may also have issues such as sample bias (Miguéis et al., 2018). This also leads to the related problem of class imbalance, where the number of samples for the classes are widely disproportionate (Cabezuelo et al., 2023; Oliveira et al., 2019; Punlumjeak et al., 2017; Sha et al., 2022).

- Increase model generalizability: There is a need to apply the models made by the researchers across

many higher education institutes (HEI) to ensure the model is accurate. Researchers are looking into applying a profit-driven approach to the model and use it on other HEIs (Alwarthan et al., 2022; Maldonado et al., 2021). Other researchers wish to apply their model to different courses (Alwarthanet al., 2022).

- Increase model accuracy by modifying model: Some researchers are attempting to increase model accuracy by modifying algorithms (Cabezuelo et al., 2023) and performing additional hyper-parameter tuning (Dileep et al., 2022). They are also looking into ensemble classification (Pratamaet al., 2021) or using a voting scheme of machine learning algorithms to increase the model accuracy (Palacios et al., 2021). Researchers are also looking into the explainablity of classifiers to understand their inner workings (Palacios et al., 2021). Researchers are also trying to incorporate their proposed models into early warning systems within the academic institute (Chen et al., 2018).

CONCLUSIONS

From all the articles that have been reviewed, the research gaps and future trends for EDM are discussed. For future work, this review can be extended into a systematic review and the number of papers may be increased. Education is one of the most important aspects in a human life and influences the rest of a person’s life. There is a need to assist universities in predicting at-risk students for the sake of students as well as the educational institute.

Acknowledgements

This research was funded by Fundamental Research Grant Scheme (FRGS), with grant number FRGS/1/2022/ICT06/MMU/03/3; and TM Research & Development Grant, MMUE/220028.

REFERENCES

[1] Abdul Bujang, S. D., Selamat, A., Krejcar, O., Mohamed, F., Cheng, L. K., Chiu, P. C., & Fujita, H. (2023). Imbalanced Classification Methods for Student Grade Prediction: A Systematic Literature Review. IEEE Access, 11, 1970–1989. https://doi.org/10.1109/ACCESS.2022.3225404

[2] Adekitan, A. I., & Salau, O. (2019). The impact of engineering students’ performance in the first three years on their graduation result

using educational data mining. Heliyon, 5(2), e01250. https://doi.org/10.1016/J.HELIYON.2019.E01250

[3] Ai, D., Zhang, T., Yu, G., & Shao, X. (2020). A Dropout Prediction Framework Combined with Ensemble Feature Selection. ACM International Conference Proceeding Series, 179–185. https://doi.org/10.1145/3395245.3396432

[4] Alboaneen, D., Almelihi, M., Alsubaie, R., Alghamdi, R., Alshehri, L., & Alharthi, R. (2022). Development of a Web-Based Prediction System for Students' Academic Performance. Data, 7(2). https://doi.org/10.3390/DATA7020021

[5] Alcaraz, R., Martinez-Rodrigo, A., Zangroniz, R., & Rieta, J. J. (2021). Early Prediction of Students at Risk of Failing a Face-to- Face Course in Power Electronic Systems. IEEE Transactions on Learning Technologies, 14(5), 590–603. https://doi.org/10.1109/TLT.2021.3118279

[6] Alija, S., Beqiri, E., Gaafar, A. S., & Hamoud, A. K. (2023). Predicting Students Performance Using Supervised Machine Learning Based on Imbalanced Dataset and Wrapper Feature Selection. Informatica, 47(1), 11–20. https://doi.org/10.31449/INF.V47I1.4519 Almasri, A., Alkhawaldeh, R. S., & Çelebi, E. (2020). Clustering-Based EMT Model for Predicting Student Performance. Arabian Journal for Science and Engineering, 45(12), 10067–10078. https://doi.org/10.1007/S13369-020-04578-4

[7] Alraddadi, S., Alseady, S., & Almotiri, S. (2021). Prediction of students academic performance utilizing hybrid teaching-learning based feature selection and machine learning models. 2021 International Conference of Women in Data Science at Taif University, WiDSTaif 2021. https://doi.org/10.1109/WIDSTAIF52235.2021.9430248

[8] Alsharari, N. M., & Alshurideh, M. T. (2021). Student retention in higher education: the role of creativity, emotional intelligence and learner autonomy. International Journal of Educational Management, 35(1), 233–247. https://doi.org/10.1108/IJEM-12-2019-0421

[9] Alwarthan, S., Aslam, N., & Khan, I. U. (2022). An Explainable Model for Identifying At-Risk Student at Higher Education. IEEEAccess, 10, 107649–107668. https://doi.org/10.1109/ACCESS.2022.3211070

[10] Alyahyan, E., & Düştegör, D. (2020). Predicting academic success in higher education: literature review and best practices. International Journal of Educational Technology in Higher Education, 17(1), 1–21. https://doi.org/10.1186/S41239-020-0177-7/TABLES/15

[11] Arif, M. A., Jahan, A., Mau, M. I., & Tummarzia, R. (2021). An Improved Prediction System of Students' Performance Using Classification model and Feature Selection Algorithm. Int. J. Advance Soft Compu. Appl, 13(1).

[12] Jam, F. A., Singh, S. K. G., Ng, B., & Aziz, N. (2018). The interactive effect of uncertainty avoidance cultural values and leadership styles on open service innovation: A look at malaysian healthcare sector. International Journal of Business and Administrative Studies, 4(5), 208-223.

[13] Assistant, M. K., Nidhi, N., Majithia, S., & Sharma, N. (2021). Predictive Model for Students' Academic Performance Using Classification and Feature Selection Techniques. Proceedings - 2021 2nd International Conference on Computational Methods in Science and Technology, ICCMST 2021, 106–111. https://doi.org/10.1109/ICCMST54943.2021.00032

[14] Awad, M., & Khanna, R. (2015). Machine Learning. Efficient Learning Machines, 1–18. https://doi.org/10.1007/978-1-4302-5990-9_1

[15] Azcona, D., Hsiao, I. H., & Smeaton, A. F. (2019). Personalizing computer science education by leveraging multimodal learning analytics. Proceedings - Frontiers in Education Conference, FIE, 2018-Octob. https://doi.org/10.1109/FIE.2018.8658596

[16] Azcona, D., & Smeaton, A. F. (2017). Targeting at-risk students using engagement and effort predictors in an introductory computer programming course. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 10474 LNCS, 361–366. https://doi.org/10.1007/978-3-319-66610-5_27/TABLES/1

[17] Bello, F. A., Kohler, J., Hinrechsen, K., Araya, V., Hidalgo, L., & Jara, J. L. (2020). Using machine learning methods to identify significant variables for the prediction of first-year Informatics Engineering students dropout. Proceedings - International Conference of the Chilean Computer Science Society, SCCC, 2020-November. https://doi.org/10.1109/SCCC51225.2020.9281280 Borrella, I., Caballero-Caballero, S., & Ponce-Cueto, E. (2022). Taking action to reduce dropout in MOOCs: Tested interventions. Computers & Education, 179, 104412. https://doi.org/10.1016/J.COMPEDU.2021.104412

[18] Burgos, C., Campanario, M. L., Peña, D. de la, Lara, J. A., Lizcano, D., & Martínez, M. A. (2018). Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout. Computers & Electrical Engineering, 66, 541–556. https://doi.org/10.1016/J.COMPELECENG.2017.03.005

[19] Cabezuelo, S., González, R., Campo, D., Barbero, R., & Mduma, N. (2023). Data Balancing Techniques for Predicting Student Dropout Using Machine Learning. Data 2023, Vol. 8, Page 49, 8(3), 49. https://doi.org/10.3390/DATA8030049

[20] Chen, Y., Johri, A., & Rangwala, H. (2018). Running out of STEM: A Comparative Study across STEM Majors of College Students At-Risk of Dropping Out Early. Proceedings of the 8th International Conference on Learning Analytics and Knowledge, 270–279. https://doi.org/10.1145/3170358

[21] Chui, K. T., Fung, D. C. L., Lytras, M. D., & Lam, T. M. (2020). Predicting at-risk university students in a virtual learning environment via a machine learning algorithm. Computers in Human Behavior, 107, 105584. https://doi.org/10.1016/J.CHB.2018.06.032

[22] Chung, J. Y., & Lee, S. (2019). Dropout early warning systems for high school students using machine learning. Children and Youth Services Review, 96, 346–353. https://doi.org/10.1016/J.CHILDYOUTH.2018.11.030

[23] Cooke, A., Smith, D., & Booth, A. (2012). Beyond PICO. Http://Dx.Doi.Org/10.1177/1049732312452938, 22(10), 1435–1443.

https://doi.org/10.1177/1049732312452938

[24] Coussement, K., Phan, M., De Caigny, A., Benoit, D. F., & Raes, A. (2020). Predicting student dropout in subscription-based online learning environments: The beneficial impact of the logit leaf model. Decision Support Systems, 135, 113325. https://doi.org/10.1016/J.DSS.2020.113325

[25] Crivei, L. M., Czibula, G., Ciubotariu, G., & Dindelegan, M. (2020). Unsupervised learning based mining of academic data sets for students' performance analysis. SACI 2020 - IEEE 14th International Symposium on Applied Computational Intelligence and Informatics, Proceedings, 11–16. https://doi.org/10.1109/SACI49304.2020.9118835

[26] Al-Shanfari, L. ., Abdullah, S. ., Fstnassi, T. ., & Al-Kharusi, S. . (2023). Instructors' Perceptions of Intelligent Tutoring Systems and Their Implications for Studying Computer Programming in Omani Higher Education Institutions. International Journal of Membrane Science and Technology, 10(2), 947-967. https://doi.org/10.15379/ijmst.v10i2.1395

[27] Czibula, G., Ciubotariu, G., Maier, M. I., & Lisei, H. (2022). IntelliDaM: A Machine Learning-Based Framework for Enhancing the Performance of Decision-Making Processes. A Case Study for Educational Data Mining. IEEE Access, 10, 80651–80666. https://doi.org/10.1109/ACCESS.2022.3195531

[28] Dileep, A. K., Bansal, A., & Cunningham, J. (2022). Early Detection of At-Risk Students in a Calculus Course. Proceedings - 2022 IEEE 46th Annual Computers, Software, and Applications Conference, COMPSAC 2022, 187 –194. https://doi.org/10.1109/COMPSAC54236.2022.00034

[29] Du, X., Yang, J., Hung, J. L., & Shelton, B. (2020). Educational data mining: a systematic review of research and emerging trends. Information Discovery and Delivery, 48(4), 225–236. https://doi.org/10.1108/IDD-09-2019-0070/FULL/PDF

[30] Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). Learning from Imbalanced Data Sets. Learning from Imbalanced Data Sets. https://doi.org/10.1007/978-3-319-98074-4

[31] Grant, M. J., & Booth, A. (2009). A typology of reviews: an analysis of 14 review types and associated methodologies. Health Information & Libraries Journal, 26(2), 91–108. https://doi.org/10.1111/J.1471-1842.2009.00848.X

[32] Guzmán-Castillo, S., Körner, F., Pantoja-García, J. I., Nieto-Ramos, L., Gómez-Charris, Y., Castro-Sarmiento, A., & Romero-Conrado, A. R. (2022). Implementation of a Predictive Information System for University Dropout Prevention. Procedia Computer Science, 198, 566–571. https://doi.org/10.1016/J.PROCS.2021.12.287

[33] Hill, F., Fulcher, D., Sie, R., & De Laat, M. (2019). Balancing Accuracy and Transparency in Early Alert Identification of Students at Risk. Proceedings of 2018 IEEE International Conference on Teaching, Assessment, and Learning for Engineering, TALE 2018, 1125–1128. https://doi.org/10.1109/TALE.2018.8615370

[34] Inyang, U. G., Eyoh, I. J., Robinson, S. A., & Udo, E. N. (2019). Visual Association Analytics Approach to Predictive Modelling of Students' Academic Performance. Modern Education and Computer Science, 12, 1–13. https://doi.org/10.5815/ijmecs.2019.12.01

[35] 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. https://doi.org/10.1016/J.DAJOUR.2023.100204

[36] Jahin, D., Emu, I. J., Akter, S., Patwary, M. J. A., Bhuiyan, M. A. S., & Miraz, M. H. (2021). A Novel Oversampling Technique to Solve Class Imbalance Problem: A Case Study of Students' Grades Evaluation. Proceedings - 2021 International Conference on Computing, Networking, Telecommunications and Engineering Sciences Applications, CoNTESA 2021, 69 –75. https://doi.org/10.1109/CONTESA52813.2021.9657151

[37] Jia, P., & Maloney, T. (2015). Using predictive modelling to identify students at risk of poor university outcomes. Higher Education, 70(1), 127–149. https://doi.org/10.1007/S10734-014-9829-7/TABLES/3

[38] Jiang, P., Liu, Z., Zhang, L., & Wang, J. (2023). Hybrid model for profit-driven churn prediction based on cost minimization and return maximization. Expert Systems with Applications, 228, 120354. https://doi.org/10.1016/J.ESWA.2023.120354

[39] Kumar, V. U., Krishna, A., Neelakanteswara, P., & Basha, C. Z. (2020). Advanced Prediction of Performance of a Student in an University using Machine Learning Techniques. Proceedings of the International Conference on Electronics and Sustainable Communication Systems, ICESC 2020, 121–126. https://doi.org/10.1109/ICESC48915.2020.9155557

[40] Kuzilek, J., Zdrahal, Z., & Fuglik, V. (2021). Student success prediction using student exam behaviour. Future Generation Computer Systems, 125, 661–671. https://doi.org/10.1016/J.FUTURE.2021.07.009

[41] Lakkaraju, H., Aguiar, E., Shan, C., Miller, D., Bhanpuri, N., Ghani, R., & Addison, K. L. (2015). A machine learning framework to identify students at risk of adverse academic outcomes. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015-August, 1909–1918. https://doi.org/10.1145/2783258.2788620

[42] Litalien, D., & Guay, F. (2015). Dropout intentions in PhD studies: A comprehensive model based on interpersonal relationships and motivational resources. Contemporary Educational Psychology, 41, 218–231. https://doi.org/10.1016/J.CEDPSYCH.2015.03.004

[43] Maldonado, S., Miranda, J., Olaya, D., Vásquez, J., & Verbeke, W. (2021). Redefining profit metrics for boosting student retention in higher education. Decision Support Systems, 143, 113493. https://doi.org/10.1016/J.DSS.2021.113493

[44] Marbouti, F., Diefes-Dux, H. A., & Madhavan, K. (2016). Models for early prediction of at-risk students in a course using standards-based grading. Computers and Education, 103, 1–15. https://doi.org/10.1016/j.compedu.2016.09.005

[45] Mariano, A. M., De Magalhães Lelis Ferreira, A. B., Santos, M. R., Castilho, M. L., & Bastos, A. C. F. L. C. (2022). Decision trees for predicting dropout in Engineering Course students in Brazil. Procedia Computer Science, 214(C), 1113–1120. https://doi.org/10.1016/J.PROCS.2022.11.285

[46] Masood, S. W., & Begum, S. A. (2022). Comparison of Resampling Techniques for Imbalanced Datasets in Student Dropout Prediction. Proceedings - 2022 IEEE Silchar Subsection Conference, SILCON 2022. https://doi.org/10.1109/SILCON55242.2022.10028915

[47] Miguéis, V. L., Freitas, A., Garcia, P. J. V., & Silva, A. (2018). Early segmentation of students according to their academic

performance: A predictive modelling approach. Decision Support Systems, 115, 36–51. https://doi.org/10.1016/J.DSS.2018.09.001 Moreno-Marcos, P. M., Muñoz-Merino, P. J., Maldonado-Mahauad, J., Pérez-Sanagustín, M., Alario-Hoyos, C., & Delgado Kloos,

[48] C. (2020). Temporal analysis for dropout prediction using self-regulated learning strategies in self-paced MOOCs. Computers & Education, 145, 103728. https://doi.org/10.1016/J.COMPEDU.2019.103728

[49] Mubarak, A. A., Cao, H., & Hezam, I. M. (2021). Deep analytic model for student dropout prediction in massive open online courses. Computers & Electrical Engineering, 93, 107271. https://doi.org/10.1016/J.COMPELECENG.2021.107271

[50] Nik Nurul Hafzan, M. Y., Safaai, D., Asiah, M., Mohd Saberi, M., & Siti Syuhaida, S. (2019). Review on Predictive Modelling Techniques for Identifying Students at Risk in University Environment. MATEC Web of Conferences, 255, 03002. https://doi.org/10.1051/matecconf/201925503002

[51] Niyogisubizo, J., Liao, L., Nziyumva, E., Murwanashyaka, E., & Nshimyumukiza, P. C. (2022). Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization. Computers and Education: Artificial Intelligence, 3, 100066. https://doi.org/10.1016/J.CAEAI.2022.100066

[52] Nuanmeesri, S., Poomhiran, L., Chopvitayakun, S., & Kadmateekarun, P. (2022). Improving Dropout Forecasting during the COVID-19 Pandemic through Feature Selection and Multilayer Perceptron Neural Network. International Journal of Information and Education Technology, 12(9), 851–857. https://doi.org/10.18178/IJIET.2022.12.9.1693

[53] Offiah, G., Ekpotu, L. P., Murphy, S., Kane, D., Gordon, A., O'Sullivan, M., Sharifuddin, S. F., Hill, A. D. K., & Condron, C. M.(2019). Evaluation of medical student retention of clinical skills following simulation training. BMC Medical Education, 19(1). https://doi.org/10.1186/S12909-019-1663-2

[54] Olaya, D., Vásquez, J., Maldonado, S., Miranda, J., & Verbeke, W. (2020). Uplift Modeling for preventing student dropout in higher education. Decision Support Systems, 134, 113320. https://doi.org/10.1016/J.DSS.2020.113320

[55] Oliveira, M. M. De, Barwaldt, R., Pias, M. R., & Espindola, D. B. (2019). Understanding the Student Dropout in Distance Learning. Proceedings - Frontiers in Education Conference, FIE, 2019-Octob. https://doi.org/10.1109/FIE43999.2019.9028433

[56] Paideya, V., & Bengesai, A. V. (2021). Predicting patterns of persistence at a South African university: a decision tree approach. International Journal of Educational Management, 35(6), 1245–1262. https://doi.org/10.1108/IJEM-04-2020-0184

[57] 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 2021, Vol. 23, Page 485, 23(4), 485. https://doi.org/10.3390/E23040485

[58] Pratama, I., Pristyanto, Y., & Prasetyaningrum, P. T. (2021). Imbalanced Class handling and Classification on Educational Dataset. ICOIACT 2021 - 4th International Conference on Information and Communications Technology: The Role of AI in Health and Social Revolution in Turbulence Era, 180–185. https://doi.org/10.1109/ICOIACT53268.2021.9563968

[59] Premalatha, N., & Sujatha, S. (2021). An Effective Ensemble Model to Predict Employment Status of Graduates in Higher Educational Institutions. 2021 4th International Conference on Electrical, Computer and Communication Technologies, ICECCT 2021. https://doi.org/10.1109/ICECCT52121.2021.9616952

[60] Punlumjeak, W., Rachburee, N., & Arunrerk, J. (2017). Big Data Analytics: Student Performance Prediction Using Feature Selection and Machine Learning on Microsoft Azure Platform. Journal of Telecommunication, Electronic and Computer Engineering (JTEC), 9(1–4), 113–117. https://jtec.utem.edu.my/jtec/article/view/1791

[61] Ranjeeth, S., Latchoumi, T. P., & Paul, P. V. (2020). A Survey on Predictive Models of Learning Analytics. Procedia Computer Science, 167, 37–46. https://doi.org/10.1016/J.PROCS.2020.03.180

[62] Rastrollo-Guerrero, J. L., Gómez-Pulido, J. A., & Durán-Domínguez, A. (2020). Analyzing and Predicting Students' Performance by Means of Machine Learning: A Review. Applied Sciences 2020, Vol. 10, Page 1042, 10(3), 1042. https://doi.org/10.3390/APP10031042

[63] Rodrigues, L. S., Dos Santos, M., Costa, I., & Moreira, M. A. L. (2022). Student Performance Prediction on Primary and Secondary Schools-A Systematic Literature Review. Procedia Computer Science, 214(C), 680–687. https://doi.org/10.1016/J.PROCS.2022.11.229

[64] Rogers, T., Colvin, C., & Chiera, B. (2014). Modest analytics: using the index method to identify students at risk of failure. Proceedings of the Fourth International Conference on Learning Analytics And Knowledge. https://doi.org/10.1145/2567574 Saluja, R., & Rai, M. (2022). Analysis of Existing ML Techniques for Students Success Prediction. PDGC 2022 - 2022 7thInternational Conference on Parallel, Distributed and Grid Computing, 507–512. https://doi.org/10.1109/PDGC56933.2022.10053236

[65] Sha, L., Rakovic, M., Das, A., Gasevic, D., & Chen, G. (2022). Leveraging Class Balancing Techniques to Alleviate Algorithmic Bias for Predictive Tasks in Education. IEEE Transactions on Learning Technologies, 15(4), 481–492. https://doi.org/10.1109/TLT.2022.3196278

[66] Shafiq, D. A., Marjani, M., Habeeb, R. A. A., & Asirvatham, D. (2022). Student Retention Using Educational Data Mining and Predictive Analytics: A Systematic Literature Review. IEEE Access, 10, 72480–72503. https://doi.org/10.1109/ACCESS.2022.3188767

[67] Shelton, B. E., Yang, J., Hung, J. L., & Du, X. (2018). Two-stage predictive modeling for identifying at-risk students. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11003 LNCS, 578–583. https://doi.org/10.1007/978-3-319-99737-7_61/TABLES/1

Tight, M. (2019). Student retention and engagement in higher education. Https://Doi.Org/10.1080/0309877X.2019.1576860, 44(5), 689–

704. https://doi.org/10.1080/0309877X.2019.1576860

[68] Verma, S., Yadav, R. K., & Kholiya, K. (2022). A Scalable Machine Learning-based Ensemble Approach to Enhance the Prediction Accuracy for Identifying Students at-Risk. International Journal of Advanced Computer Science and Applications, 13(8), 185–192. https://doi.org/10.14569/IJACSA.2022.0130822

[69] Zhu, R., Guo, Y., & Xue, J. H. (2020). Adjusting the imbalance ratio by the dimensionality of imbalanced data. Pattern Recognition Letters, 133, 217–223. https://doi.org/10.1016/J.PATREC.2020.03.004

DOI: https://doi.org/10.15379/ijmst.v10i3.1841