Education and Information Technologies (2023) 28:9235–9266
https://doi.org/10.1007/s10639-022-11571-x

On predicting school dropouts in Egypt: A machine learning approach

Kamal Samy Selim[1] · **Sahar Saeed Rezk**[1]

Received: 21 September 2022 / Accepted: 28 December 2022 / Published online: 12 January 2023

Abstract

Compulsory school-dropout is a serious problem affecting not only the education systems, but also the developmental progress of any country as a whole. Identifying the risk of dropping out, and characterizing its main determinants, could help the decision-makers to draw eradicating policies for this persisting problem and reducing its social and economic negativities over time. Based on a substantially imbalanced Egyptian survey dataset, this paper aims to develop a Logistic classifier capable of early predicting students at-risk of dropping out. Training any classifier with an imbalanced dataset, usually weaken its performance especially when it comes to false negative classification. Due to this fact, an extensive comparative analysis is conducted to investigate a variety of resampling techniques. More specifically, based on eight under-sampling techniques and four over-sampling ones, and their mutually exclusive mixed pairs, forty-five resampling experiments on the dataset are conducted to build the best possible Logistic classifier. The main contribution of this paper is to provide an explicit predictive model for school dropouts in Egypt which could be employed for identifying vulnerable students who are continuously feeding this chronic problem. The key factors of vulnerability the suggested classifier identified are student chronic diseases, co-educational, parents' illiteracy, educational performance, and teacher caring. These factors are matching with those found by many of the research previously conducted in similar countries. Accordingly, educational authorities could confidently monitor these factors and tailor suitable actions for early intervention.

Keywords School-Dropout · Classification · Logistic Classifier · Class Imbalance · Imbalance Learning

✉ Sahar Saeed Rezk
saharsaeed@cu.edu.eg

Kamal Samy Selim
kamalselim@feps.edu.eg

[1] Department of Socio-Computing, Faculty of Economics and Political Science, Cairo University, Giza, Egypt

1 Introduction

Education is on the top of basic human rights that grantees children and adolescents to develop and acquire the knowledge and skills required to realize their full potential and participate actively in their society. The fourth goal of Sustainable Development asks for all girls and boys to have access to free, equitable, and high-quality primary and secondary education by 2030, resulting in relevant and effective learning outcomes. To accomplish this goal, it is critical that each child completes his/her education without dropping out (UNICEF, 2017).

School-dropout is defined as leaving school before completing an education cycle or program that has already begun. Students at-risk of dropping out are those registered in any mandatory or post-mandatory program but are exhibiting risk factors or symptoms that indicate they may drop out (UNICEF, 2017).

Building a system to early predict students at-risk of school-dropout or being able to characterize the main determinants of this problem in advance could help reducing its negative social and economic implications. It also might help to provide policymakers with guidance so as to eradicate the causes of this social behavior over years.

Many of the existing studies have been developed to look at the factors influencing school-dropout in Egypt. Even though, none of them provides an explicit classification model that can be used as an early warning system for predicting this chronic problem. To fill this gap in the Egyptian relevant work, this paper aims at developing a well-performing Logistic classifier capable of early predicting at-risk students based on a substantially imbalanced Egyptian survey dataset collected in 2014.

Generally, there are two types of errors in classification models, called Type I and Type II errors. Type I error (also known as false positive (Liang et al., 2018)) emerges, for example, when the classifier mistakenly labels a student who is actually not likely to dropout schooling as being at-risk of doing so. On the other hand, Type II error (often called false negative (Gonzalez-Abril et al., 2017)) arises when the classifier incorrectly labels a student who is likely to dropout as being in the class of not dropping out. In this study, Type II error is considered significantly more important than Type I error. This is because it is believed that a higher Type II error rate results in higher social costs and puts students' education potentials at danger.

In fact, the class-imbalance is one of the most typical issues that leads to both Type I and Type II errors. In many applications, it represents a tricky problem to solve in classification tasks, especially with those having a binary class setup such as the problem of school-dropout at hand. It could be said that the fundamental difficulty with imbalance learning is that it leads to minority class misclassification resulting in inaccurate classifiers (Elreedy & Atiya, 2019). In general, when class frequencies are imbalanced, traditional algorithms, such as Logistic Regression (LR), may perform poorly for the unseen instances of the minority class. This is owing to the fact that the majority class has a significant impact on the model preventing it from reliably classifying instances belonging to the minority class (Amin et al., 2016; Mohammed, 2020).

Goel et al. (2013) elucidate that sampling, cost, kernel, and active learning-based algorithms have all been developed to address the learning challenges of imbalanced datasets. The current study emphasizes on the sampling strategies. In other words, this paper uses an Egyptian school-dropout imbalanced dataset to investigate how a good determination of a resampling technique could significantly improve the Logistic classifier performance. More specifically, an extensive comparative analysis of several resampling techniques is conducted to determine the best one to integrate with the Logistic classifier so as to improve its power based on three of the common performance measures.

After this introductory section, the paper is organized as follows. In section two, the previous works on the problem of school-dropout are reviewed. In section three, an exposition of the research methodology is provided, involving LR and resampling techniques. The implementation phases are explained in-depth in section four, followed by a thorough discussion of the experimental results in section five. Section six is devoted to the detailed explanation of the resulting classification model from an educational point of view. Finally, section seven concludes the study and makes some recommendations for future research.

2 School-dropout problem: A review of literature

This paper is interested in the problem of school-dropout during the level of basic education. To put it another way, the main purpose of the study is to develop a classification model having an ability to early predict students at-risk of dropping out during the compulsory school years.[1] To highlight the limitations of the previous works, Table 1 summarize the relevant literature on school-dropout modeling.

The above review reveals that firstly, LR is one of the main algorithms that are recommended for school-dropout classification problems. Some of the above-mentioned studies demonstrate that LR can outperform other classification algorithms in terms of overall accuracy of early school-dropout prediction. Secondly, previous works for the problem of dropout in Egypt are scarce and have some shortcomings. To name it, no study is undertaken for predicting students at-risk of dropping out during the basic education. Also, the majority of these studies investigate the factors that influence dropping out, but none of them targets an explicit model that can be used as part of an early warning system for this chronic problem. Accordingly, the current study aims to fill this gap by developing a Logistic classifier that can be utilized for this purpose. It also aims at investigating possible remedies of the biasing effects resulting from the class-imbalance problem featuring many of the real-world datasets. Methodologically, this is achieved by implementing under- and over-sampling techniques to improve the constructed model's classification performance.

[1] In Egypt, both the primary and preparatory stages of schooling make up together the compulsory level of basic education.

Table 1 Overview of literature in school-dropout modeling – created by researchers

Reference	Focus and Main Results	Methodology
(Lloyd et al., 2001)	The study investigates the empirical association between the school level of quality and the probability of dropping out during the preparatory and secondary school levels in Egypt. Findings of the study show that school quality is significantly affecting school levels completion	Discrete-time hazard model
(Safaa & El-Daw, 2001)	In both Yemen and Egypt, the researchers look at how poverty affects human capital development. According to the study's findings for Egypt, children from low-income households are less likely to attend school, and if they do, they are more vulnerable to drop out and engage in child labour	Principal component analysis
(Suliman & El-kogali, 2002)	With a particular focus on gender disparities, the study explores the factors influencing children's access to and completion of basic education in Egypt. The findings indicate that Egypt is still facing problems in achieving universal basic education, particularly for girls. Moreover, findings indicate that household level of wealth, parents' education, and the cost of schooling per pupil are the most important predictors of children's dropout	Multinomial logistic model
(Hanushek et al., 2006)	This study looks into the relationship between school quality and the likelihood of completing the primary school level in Egypt. The findings indicate that students who attend a low-quality school are substantially less likely to stay in school than students who attend a high-quality school	Mathematical optimization model
(Assaad, 2010)	The study examines the impact of girls' work on their school attendance. Findings reveal that girls who do a lot of household labour have lower school attendance rates. Policies forbidding children from working in the labour market will have little impact on girls' schooling, whereas efforts to alleviate the pressure of domestic labour, such as improving water and sanitation facilities, will have a better chance of success	Bivariate probit approach
(Quadri & Kalyankar, 2010)	The study illustrates how Decision Trees may be used to early predict a student's likelihood of dropping out. The list of students who are anticipated to be at-risk for dropping out of college is then sent to teachers and management for direct or indirect assistance	Decision Trees
(Mali et al., 2012)	This study is being carried out to ascertain the severity of the school-dropout issue and the contributing factors in Maharashtra, India. In this context, the study adopts a descriptive cross-sectional design with a nationwide focus to identify school dropouts and their socio-demographic characteristics	Chi-square and Fisher's exact test

Table 1 (continued)

Reference	Focus and Main Results	Methodology
(Rahaman & Das, 2018)	The study investigates how socioeconomic, demographic, and school-related factors may affect dropout rates in Manipur, India. Based on the study's main results, socioeconomic and school-related factors were the main causes of children leaving school	LR
(Hasan, 2019)	The main objective of the study is to compare the linear discriminant model's prediction accuracy with that of LR using a real dataset of female students' school-dropout rates in Bangladesh	LR and Linear Discriminant Analysis
(Mduma et al., 2019)	With Tanzania taken as a case study, this research aims at identifying an algorithmic suggestion to the researchers in the topic of predicting school dropouts. The main findings specify LR and Multilayer Perceptron as the best performing algorithms	Machine Learning (ML) Techniques
(Radwan, 2019)	The study seeks to figure out what causes females to drop out of school and what the consequences are in rural Egypt. According to the study, the most major socio-cultural determinants for girls' school-dropout are a preference for male education when it comes to scarcity of income, and the death of one of the parents or both. While the most important economic drivers for the problem are limited family income and having many children	Simple statistical correlation coefficients
(Sarra et al., 2019)	The study aims to assess the usefulness of a specific latent class model for the identification of Italian students at-risk of university dropping out	Bayesian Profile Regression
(Timbal, 2019)	In this study, a data mining technique is used to build a rule-based classifier. Three rules are generated from the decision tree based on the findings and used directly as a basis to predict whom enrolled students will drop out	Decision Trees
(Kabathova & Drlik, 2021)	The essential goal of the study is to evaluate the effectiveness of some ML algorithms in terms of drop-out prediction. The findings indicate that LR, among other techniques, might be successfully used to analyze a small sample of academic data	ML Techniques
(Mnyawami et al., 2022)	The study uses a number of frequently used ML techniques to predict secondary school dropouts in Tanzania. The main findings show that Decision Trees offer an accurate list of factors that contribute to student dropout	ML Techniques

3 Methodology

This section provides an overview of one of the most common classification algorithms which is LR, as well as an exposition of the main resampling techniques designed to deal with class-imbalance problems. In addition, a brief description of the metrics that could be used to examine the classifier performance is presented. But before going any further, a technical illustration of classification setup is needed so that this review could be formally followed.

Given a dataset $D = [(X_i, y_i); i = 1, 2, \dots, n]$ where for each instance i the vector $X_i = (x_{i1}, x_{i2}, \dots, x_{im})$ is a realization of m finite variables $x_j; j = 1, 2, \dots, m$ representing the set of categorical and/or numerical attributes of concern, and y_i is the class value. In this study, it is presupposed that each instance belongs to only one of two classes (i.e., $y_i \in \{0, 1\}$).

Following the completion of sufficient training/learning, the classification task is to develop a function that is able to map the inputs of vector X_i into an output y_i by using such supervised methods that are referred to as classification algorithms. A classification model or classifier is usually the name given to the resulting function (f) which enables the discovery of hidden links between the target class and independent/explanatory attributes. Once the classifier is developed, its performance could be estimated using one of the evaluation metrics (Avon, 2016; Berrar, 2018).

3.1 Logistic regression (LR)

LR analysis is commonly employed in order to investigate the association between a categorical dependent variable and a set of independent/explanatory variables. In binary classification problems, it is often utilized to model the likelihood of a specific class or event occurring given a set of some predictors as presented by the following equation.

$$p\left(y_i = 1 \middle| X_i\right) = \frac{\exp(\beta_0 + \beta^T X_i)}{1 + \exp(\beta_0 + \beta^T X_i)}$$

where $y_i = 1$ when the event occurs versus $y_i = 0$ when it does not (e.g., student drops out schooling versus she/he does not). β_0 is the intercept term and β^T is the transpose of regression coefficients vector. After some mathematical transformations, the LR model employs the natural logarithm of the odds as a regression function of the independent variables. This takes the form of the following equation.

$$ln\left(\frac{p(y_i = 1|X_i)}{1 - p(y_i = 1|X_i)}\right) = \beta_0 + \beta^T X_i$$

As a well-known classification algorithm, the main advantages of LR are as follows. Firstly, the produced Logistic model is easily to be interpreted and understood. This feature is widely desired in applied research disciplines, especially for studies assisting the policymakers in taking decisions. Secondly, it is generally a flexible technique when it comes to analyzing mixed datasets (Peng et al., 2002; Tansey et al., 1996). Third, LR is also considered as an effective model for feature reduction by embedding additional constraints on the parameter space of the optimization problem. In order to do that, the model adds regularization penalties which is a crucial task to prevent overfitting, especially if there are a small number of data instances but many features. The LR model can be regularized in a variety of ways. When data contain irrelevant features, two popular methods have been shown to have a good performance, namely LASSO (L1) and RIDGE (L2). L1 uses a penalty term in the model to reassemble the absolute values of the features' coeffecients into the smallest sum possible, while L2 seeks to minimize the sum of coeffecients' squares (Kristoffersen & Hernandez, 2021). In the empirical part, this study investigates the impact of both penalty options for seeking the best predictive model.

3.2 Resampling techniques

An imbalanced dataset is one that has an unequal distribution of class frequencies. The annoying magnitude of such imbalances is not universally agreed upon. Some researchers examine data where one class is few times smaller than others, while others look at more drastic imbalance ratios (Napierala & Stefanowski, 2012). Other studies such as (Kraiem et al., 2021) assume that a dataset is imbalanced when the ratio of majority to minority instances is more than 2:1. However anywise this critical ratio could be, the class-imbalance problem has generally a significant impact on the performance of ML classification techniques. In various disciplines, resampling represents an effective strategy for dealing with this problem so as to achieve reliable learning from imbalanced datasets (Amin et al., 2016). Resampling methods focus on balancing the distribution of instances belonging to minority and majority classes regardless of what their true distribution could be. Nonetheless, it is confirmed that balanced datasets help classifiers learn more accurately than imbalanced ones (Goel et al., 2013).

Overall, under-, over-sampling, and a combination of both (often called hybrid-sampling) are the three broad categories of data resampling in ML. Shamsudin et al. (2020) suggest that it is better to apply hybrid-sampling because it handles some of the individual techniques' problems such as the loss of information in case of under-sampling, and the overfitting in case of over-sampling. In the following subsections both under- and over-sampling techniques are briefly elucidated.

3.2.1 Under-sampling

Under-sampling is the procedure of reducing the number of majority class instances either at random or by applying specific algorithms. By randomly eliminating some of the majority class instances, which is known as Random Under-Sampling (RUS), loss of important information could be resulted in (Yi et al., 2022). The most popular under-sampling algorithms, on the other hand, include Edited Nearest Neighbours Rule (ENN) (Wilson, 1972), Tomek-Links (Tomek, 1976), One-Sided

Selection (OSS) (Kubat & Matwin, 1997), Neighborhood Cleaning Rule (NCL) (Laurikkala, 2001), and NearMiss (Mani & Zhang, 2003).

ENN was suggested by Wilson (1972) to discard ambiguous and noisy instances in a dataset. The process begins by determining the k-nearest neighbours of each instance of the majority class (e.g., $k = 3$), then majority class instances which are misclassified by their k-nearest neighbours are removed.

Tomek-Links was firstly proposed by Tomek (1976) to under-sample the Tomek-links which are the pairs of instances that are nearest neighbours to one another, but belong to different classes. In this manner, if two instances create a Tomek-link, it means that either one of them is noise or both are close to a decision border. As a result, this method can be used to clear up undesired overlaps across classes by eliminating all Tomek-links until all minimally distant nearest neighbour pairs are of the same class. Consequently, this strategy can be employed in the training set to produce well-defined class clusters, resulting in precise classification rules and improved classification performance (He & Garcia, 2009).

OSS was initially described by Kubat and Matwin (1997). In this algorithm, the process starts by utilizing the k-Nearest Neighbours algorithm to classify all the majority class instances, typically with $k = 1$. Then, all the minority class instances as well as the misclassified instances belonging to the majority class are selected in order to find the Tomek links among them. Finally, the majority class instances involved in the Tomek links are removed (Loyola-González et al., 2016).

NCL, as suggested by Laurikkala (2001), modifies the ENN method by increasing the role of data cleaning as follows. First, NCL removes majority class instances which are misclassified by their k-nearest neighbors. Second, the neighbours of each minority class instance are identified and the ones belonging to the majority class are removed.

NearMiss is a term used by Mani and Zhang (2003) to describe a group of under-sampling techniques. NearMiss-1, NearMiss-2, and NearMiss-3 are the three variants of this technique. NearMiss-1 excludes majority class instances with the smallest average distance to the three closest minority class instances. NearMiss-2 eliminates majority class instances with the smallest average distance to the three furthest minority class instances. NearMiss-3 removes, for each minority class instance, a predetermined number (three by default) of the closest majority class instances to make sure that every minority instance is surrounded by some majority instances (He & Garcia, 2009).

3.2.2 Over-sampling

Over-sampling can be achieved by producing new instances or repeating existing ones of minority class. Similar to under-sampling, over-sampling can be done at random or by employing particular techniques. Random Over-Sampling (ROS) duplicates instances of the minority class, which can lead to overfitting in some algorithms (Yi et al., 2022). While, Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002), Over-sampling using Adaptive Synthetic

(ADASYN) (He et al., 2008), and Borderline-SMOTE (Nguyen et al., 2011) are some of the most prominent over-sampling techniques.

SMOTE was proposed by Chawla et al. (2002) to generate synthetic instances by comparing the feature spaces of existing minority class instances. The main process of this method is as follows. First, for each minority class instance, say i, a predetermined number of the closest neighbours are found. Second, to produce a new synthetic instance, a randomly determined neighbor i^* is chosen, then the difference of the two corresponding feature vectors is multiplied by a random number γ between [0, 1] and then added to the original vector X_i (He & Garcia, 2009) as clarified by the following equation: $X_{synthetic\_i} = X_i + \left(X_{i^*} - X_i\right) * \gamma$

ADASYN, according to He et al. (2008), is an adaptive version of the original SMOTE. The fundamental idea behind this algorithm is to use a density distribution as a criterion for deciding how many synthetic instances are required to be generated for each minority instance. This is achieved by adaptively altering the weights of different minority instances to adjust for the skewed class distributions. That is, it produces more instances in regions of the feature space with a low density of minority class instances, and fewer or none in regions with a high density.

Borderline-SMOTE, as described by Nguyen et al. (2011), is an updated version of the initial SMOTE, in which borderline minority instances that are most likely to be misclassified are identified and exploited to generate new synthetic instances. The algorithm's basic procedure entails finding minority class instances that have majority class neighbours more than minority class neighbours and exploiting them for oversampling using SMOTE.

3.2.3 Resampling techniques in educational applications

Based on a search in the Scopus database, Figure 5[2] (see Appendix 1) illustrates the number of articles and conference papers published at the top 10 publication domains over the previous fifteen years (2008–2022). Whereas Figure 6 depicts the top 10 countries of publications in the areas of class-imbalance learning. Compared to the enormous number of publications in imbalance-learning (more than 10,000 are found), it could be concluded from these figures that there is a scarcity in terms of applying the resampling methods in social fields (e.g., education), especially in the developing countries.

Resampling techniques are used by some researchers to address educational issues along with enhancing the performance of the learning algorithms. Some of these studies are summarized in Table 2. Detailed review of the resampling techniques and their use in other domains could be found at (Haixiang et al., 2017) and (Kraiem et al., 2021).

[2] The search rule that is used for data extraction is: (((resampling) OR (imbalance AND learning) OR (class AND imbalance)) AND ((data AND mining) OR (machine AND learning) OR (classification))). Note that search process is within article's title, abstract, and keywords. Also, it is limited to the period from 2008 to 2022 and to only articles and conference papers.

Table 2 Resampling techniques in educational application– created by researchers

Study	Focus and Main Objectives
(Thai-Nghe et al., 2009)	The objective of this research is to use SMOTE over-sampling to address the class-imbalance and to improve the overall prediction of student performance in Can-Tho University, Vietnam. The key findings demonstrate that, in contrast to the baseline classifiers, the classifiers performance is improved after applying the resampling technique
(Koutina & Kermanidis, 2011)	The authors of this study compare the efficacy of various ML algorithms in predicting the performance of postgraduate students. They enhance the algorithms' overall performance using some resampling techniques applied in WEKA
(Chau & Phung, 2013)	In an academic credit system, the study presents a method for classifying imbalanced educational data with numerous classes based on student performance using a hybrid resampling algorithm and Random Forest
(Rashu et al., 2014)	In this study, various data mining techniques are employed to predict students' final course grades. By using SMOTE, ROS, and RUS to address the issue of a class-imbalance dataset, the prediction accuracy is increased according to the experimental results
(Yehuala, 2015)	In this research, Decision Trees and Naïve Bayes classifiers are used to predict student success and failure in the Debre_Markos university, Ethiopia. The resampling techniques are used to improve the performance in terms of the prediction of positive instances
(Radwan & Cataltepe, 2017)	The main goal of this study is to use ML techniques to predict students' performance on two educational data sets. The authors balance the training datasets by employing the SMOTE over-sampling method in order to improve the models' performance and address the class imbalance problem
(Agustianto & Destarianto, 2019)	In order to produce the most accurate Decision Tree C4.5 for predicting academic success and failure, the study focuses on balancing the student datasets utilizing the NCL under-sampling technique. The results demonstrate that NCL is able to improve the accuracy of the model predictions
(Orooji & Chen, 2019)	The essential purpose of the study is to use ML techniques to predict high school-dropout in Louisiana, USA. The case weighting and cost-sensitive learning are employed so as to improve the prediction performance on the minority class
(Ghorbani & Ghousi, 2020)	In this study, the class-imbalance problem is addressed by comparing five resampling approaches, including Borderline-SMOTE, ROS, SMOTE, SMOTE-ENN, and SMOTE-Tomek, while predicting student academic performance at two educational institutions of Iran and Portugal
(Ratih et al., 2022)	The LR and SMOTE over-sampling are used to classify student satisfaction in one of the educational institutions in Surabaya, Indonesia. According to the study's findings, the SMOTE combined with LR is more accurate and has a lower error rate than the baseline LR that is applied to the original imbalanced dataset

Table 3 Confusion matrix for binary classification

		Predicted Class	
		1	0
Actual Class	1	True Positive (TP)	False Negative (FN)
	0	False Positive (FP)	True Negative (TN)

3.3 Model evaluation metrics

In the binary classification process, instances of each class, whether being classified correctly or incorrectly, could be counted and arranged in what is known as confusion matrix representing the four possible outcomes. As illustrated in Table 3, the correctly classified instances appear on the two cells of the matrix main diagonal, whereas the off-diagonal cells reveal the numbers of instances that have been misclassified.

Based on the confusion matrix, a variety of regularly used metrics for evaluating a classifier's performance are suggested with varying evaluation emphases, including overall accuracy, precision, sensitivity/recall, specificity, Type I error, Type II error, F-score, and the area under the ROC curve (AUC). Detailed information about these measures could be found in (Maimon & Rokach, 2015).

For most of modeling techniques, the accuracy is the commonly used evaluation metric. When dealing with imbalanced datasets, however, it is not a good metric to use (Goel et al., 2013). Other evaluation metrics such as sensitivity/recall, precision, ROC-AUC, and F-score are being more widely used in such situations. Consequently, the present study employs ROC-AUC, F-score, and Type II error for comparison purposes.

Receiver Operating Characteristic (ROC-curve) is a two-dimensional representation of the trade-off between true positive (i.e., sensitivity) and false positive (i.e., type I error) rates. AUC is the area beneath the ROC-curve that measures a classifier's ability to discriminate between classes. It is the probability that the classifier will prioritize a randomly selected positive instance higher than a randomly selected negative instance. For more details and representations see (Hsu et al., 2015). F-score is the harmonized mean of both precision and recall, and it could be calculated as follow: $\left[F - score = \frac{2*TP}{FP+FN+2*TP}\right]$. Finally, Type II error is the false negative rate that is calculated using the following formula: $\left[Type\ II\ error = \frac{FN}{TP+FN}\right]$.

4 Modelling implementation

Figure 1 represents the conceptual architecture of the proposed school-dropout classification model. As illustrated, the model has two main phases. The first involves data targeting, data manipulation, and finally model development. The second, on the other hand, demonstrates how the proposed model could be used as a warning system by school authorities to early identify students who are at-risk of dropping out. The following subsections discuss the building blocks of this process in detail.

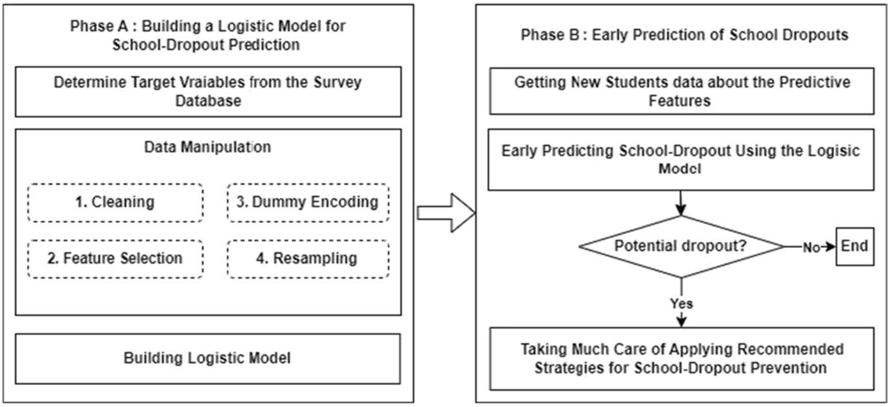

Fig. 1 Conceptual presentation of school-dropout predictive model – created by researchers

4.1 Data exploration

For the purpose of this study, a subset of the publicly available database of the Survey of Young People in Egypt (SYPE) is extracted. SYPE is a longitudinal survey that is nationally representative and offers a unique perspective on the needs and ambitions of young people across time. It mostly targets developing evidence-based programmes and policies for the sake of improving the potentials and wellbeing of the Egyptian youth. Overall, 10,916 of the youth (aged 13–35) were interviewed in the 2014 round. This poll provides gender-specific data on civic engagement, health, education, and employment. The survey's main findings, respondents' characteristics, and relevant policy implications are reported in (Population Council, 2015). In this study, some variables of the database are extracted as is, while others are composed based on some of the pre-existing attributes. In general, 18 independent variables are chosen from the survey, and they are divided into four groups based on how they relate to schools, families, students, and educational performance. Description of these variables, their associated domains, and their classes distributions among both dropouts and non-dropouts are presented in Table 9 (see Appendix A).

4.2 Data manipulation

As one of the most important steps in any data mining process, data manipulation often consumes the majority of the time and effort. It is mostly used to transform raw data into a clean dataset for the sake of improving the efficiency of data analysis. Overall, manipulation process is conducted on the selected variables as follows. First, to deal with missing values in some of the independent attributes, some instances are rejected. Second, dummy encoding is utilized to convert categorical to binary features. It is employed to only two variables, including number of siblings and individual's birth order. Third, only relevant attributes correlated to class label are selected for modelling based on the Chi-Square (χ^2) test. Attributes with 95 percent or higher confidence are considered having significant association with

Table 4 Results of Chi-squared test of correlation

Variable	Chi-Score	Variables	Chi-Score
Gender	0.25	Nursery	64.75**
Place of residence	0.95	Punishment	56.55**
Parents illiteracy	167.73**	Co-educational	170.91**
Father's employment status	0.02	Shifts	4.44**
Mother's employment status	0.00	Equal treatment	8.44**
Poverty	145.32**	Private tutoring	0.00
Siblings	81.98**	Class-fail	68.10**
Birth order	25.29**	Year repetition	27.04**
Chronic diseases	6.60**	Teacher caring	20.24**

Level of significance: ** $p < 0.05$

the class label. The five insignificant features namely, gender, place of residence, father's employment status, mother's employment status, and private tutoring are eliminated in the subsequent modelling steps. Table 4 summarizes the test results.

Accordingly, the workable dataset is consisted of 3154 records/instances each of them has one response value about school dropping out and 15 values about the selected classification features, including the dummy ones. Further, the distribution of class instances reveals that the target group (i.e., school dropouts) approximately accounts only for 19% of the total number of cases in the dataset. Since the ratio of majority to minority instances is greater than 2:1, the dataset at hand has a class-imbalance problem.

4.3 Experimental setup

The experimental work for this study is designed to examine the performance of the Logistic classifier before and after applying resampling techniques, by training the model with 10-fold stratified cross validation on 80% of the data and testing on 20% to prevent data leakage. An m-fold cross validation process works as follows. At the beginning, the original training dataset is randomly divided into m equal-sized subsets. A single subset of the m subsets is used for validation, and the leftover $(m - 1)$ subsets are used for training. Accordingly, in the experiments, the process is repeated m times, and the average results of the m validations are reported. A crucial part of the cross validation is that resampling techniques are only used on the folds holding the training sets in each iteration; the validation sets are not resampled. The goal is to avoid the problem of overfitting and make sure that the induced classifier offering adequate metric values that may be applied to real instances that are distinct from the training set (Kraiem et al., 2021). Further, the testing dataset partition is also excluded from the resampling procedure for the same reasons.

As the current study aims at building the best possible Logistic classifier for the class imbalanced dataset at hand, an extensive comparative analysis is conducted to investigate the classifier's performance under a variety of resampling techniques. More specifically, based on the eight under-sampling and the four over-sampling techniques previously reviewed in subSect. 3.2, as well as all their mutually exclusive combined pairs, forty-five experiments are conducted.

Computationally, the experiments are implemented by employing spyder (3.8) as a working platform to code Python 3 programming language. Moreover, various packages including matplotlib (3.3.2), pandas (1.1.3), scikit-learn (1.0.2), Imblearn (0.0), and numpy (1.19.2) are utilized. For the whole experiments two parameters have to be determined; the targeted resampling ratio of majority versus minority instances in the training datasets and the value of k in the k -nearest neighbours-based resampling techniques.

For experiments implementing a stand-alone resampling technique whether under-sampling or over-sampling, the targeted balancing ratio is set to 1:1 i.e., to equalize the number of instances in the majority and minority classes. When using hybrid resampling, the over-sampling ratio is set to 1:2 and the under-sampling ratio is set to 1:1. This means that synthetic minority instances are initially generated until the ratio equals 1:2, then, majority instances are eliminated until both classes have the same number of instances. These values were chosen after performing some prior investigations in order to avoid overfitting in the case of over-sampling and information loss in the case of under-sampling. Further, working with multiplicity of values for this ratio is avoided as it would have resulted in an unmanageable number of findings because of the study's large number of experiments and resampling methods.

For the number of nearest neighbours (k) needed for each resampling technique, the default packages' values are used. Accordingly, the values implemented are as follows. $k = 3$ for ENN, NCL, NearMiss-1, NearMiss-2, and NearMiss-3, and $k = 5$ for SMOTE, ADASYN, and Borderline-SMOTE. Whereas these values will be hyper-tuned later for the best techniques.

5 Experiments and results

As the empirical work of this study is extensive, this whole section is devoted to the technical aspects of classifier construction. In subSect. 5.1, the performance of the Logistic classifier prior to data resampling is presented, subSect. 5.2 discusses the performance after data resampling, while subSect. 5.3 examines how well a Logistic classifier performs when using hybrid techniques that combine under- and over-sampling. The outcomes of hyper-parameter tuning are illustrated in subSect. 5.4. Finally, the final model is explained in subSect. 5.5. Meanwhile, section six discusses all of these findings from the implementation and educational policy-making perspectives.

5.1 Performance of logistic classifier before resampling

Before using any of the resampling techniques, the average scores of the tenfold cross-validation of the Logistic classifier shows that the AUC is 0.77 which is deemed satisfactory when compared to the performance of social classifications in

Table 5 Performance of Logistic classifier with resampling techniques – created by researchers

Category	Performance	Average scores of stratified tenfold cross validation			Scores of model performance on testing set		
	Technique	AUC	Type II Error	F-score	AUC	Type II Error	F-score
Under-sampling	RUS	0.769	0.283	0.463	**0.776**	0.286	0.512
	Tomek-Links	0.771	0.812	0.276	0.774	0.880	0.196
	ENN	0.771	0.423	**0.467**	**0.776**	0.421	**0.515**
	OSS	0.771	0.812	0.276	0.774	0.880	0.195
	NCL	0.771	0.518	0.442	0.773	0.541	0.467
	NearMiss-1	0.517	0.389	0.292	0.518	0.429	0.313
	NearMiss-2	0.576	**0.261**	0.323	0.600	**0.256**	0.369
	NearMiss-3	0.752	0.307	0.463	0.750	0.308	0.493
Over-sampling	ROS	**0.771**	**0.279**	**0.461**	**0.775**	0.308	**0.496**
	SMOTE	0.749	0.307	0.440	0.749	**0.286**	0.487
	ADASYN	0.740	0.300	0.432	0.745	0.301	0.483
	Borderline-SMOTE	0.745	0.300	0.434	0.752	0.293	0.481

general. The F-score is 0.27, and Type II error is 0.82. On the other side, the model's performance on unseen data is slightly worse with an AUC of 0.77, an F-score of 0.19, and a Type II error of 0.89. Clearly, these results reflect poor performance of the Logistic model when it comes to the classification of school dropouts dataset under this basic scenario.

5.2 Performance of logistic classifier with resampling

The evaluation metrics for the Logistic classifier with all of the aforementioned resampling approaches in subSect. 3.2 are shown in Table 5. For every applied technique, the best performance in each evaluation metric is underlined.

In all considered under-sampling techniques, both validation and testing scores show that the logistic model has almost the same AUC values, with the exception of the NearMiss versions. Moreover, *NearMiss-2* is the technique produces the lowest Type II error followed by *RUS*. Last but not least, *ENN* performs better in terms of F-score followed by *NearMiss-3* in case of validation assessment and *RUS* in case of assessment with testing dataset.

As for over-sampling, Table 5 shows that *ROS* generates higher values of AUC and F-score in both validation and testing evaluation processes. Further, it produces the lowest Type II error with cross-validation assessment, whereas *SMOTE* is the one produces the lowest error in terms of testing assessment.

These findings show that no conclusive 'best model' is reached yet due to the fact that performance of each resampling method implemented is conditioned on feature-related characteristics under which it performs better. However, because the study's major purpose is to construct a classification model that can be used to early predict school-dropout, it is thought that model performance with unseen dataset is deemed to be more essential. In this scenario, *NearMiss-2* and

SMOTE may be the best techniques, if the primary goal is to only reduce Type II error. Nevertheless, these experimental comparisons are further extended to include applying hybrid combinations of both under- and over-sampling techniques to get a more general conclusion.

5.3 Performance of logistic classifier with hybrids combining under- and over-sampling

Table 6 illustrates the performance metrics of Logistic classifier with hybrids combining under- and over-sampling techniques. The overall results, based on both cross-validation and testing scores, show that combining *ROS* with any of the under-sampling techniques except the first two versions of NearMiss gives the best value of AUC. Also, combining *ROS* and *NearMiss-3* results in the classifier having the lowest and consequently the best Type II error for validation processes, while the combination of *ADASYN* and *NearMiss-3* is the best with the testing assessment followed by *ROS* and *NearMiss-3*. The findings also expose that applying *ROS* with *ENN* provides the best F-score in terms of validation processes compared to the combination of *ROS* and *NCL* in case of assessment with unseen instances.

In general, it could be said that combining under- and over-sampling techniques enhances the overall outcomes, especially in terms of Type II error and F-score. For the cross-validation scores, the Type II error drops on average from 0.82 to 0.11, and the F-score rises on average from 0.27 to 0.47 when compared to the performance of the Logistic classifier using the original dataset with class-imbalance. Moreover, the performance with the unseen instances is also improved as the Type II error decreases from 0.89 to 0.10, and the F-score increases from 0.19 to 0.52.

Overall, the combination of the *ROS* with *NearMiss-3* is believed to be relatively the best for fitting the final Logistic model on the dataset at hand. This is because it yields the lowest Type II error in validation processes and ranks second in the testing assessment. Furthermore, it is one of the combinations that produce high values of both AUC and F-score in both evaluation phases. Figure 2 illustrates the tenfold cross validation scores of the Logistic classifier before sampling and after applying the hybrid of *ROS* and *NearMiss-3*.

It is worth noting that Kraiem et al. (2021) recommended the combination between SMOTE and Tomek-Links as the most suitable method for high imbalance ratio when the interest is in the recall measure (or Type II error). Based on the above results, the best identified combination outperforms the results of Kraiem et al. (2021) for the dataset under consideration.

5.4 Performance of logistic classifier with hyper-parameter optimization

Parameters of the chosen hybrid resampling method (i.e., *ROS with NearMiss-3*) along with the Logistic classifier are hyper-tuned with Grid-Search for the sake of improving the performance based on F-Score. The tested parameters and the best identified values are illustrated in Table 7. Although, these values improve

Table 6 Performance of Logistic classifier with Hybrids combining resampling techniques – created by researchers

Combined-Techniques		Average scores of stratified tenfold cross validation			Scores of model performance on testing set		
Over-Sampling	Under-Sampling	AUC	Type II Error	F-score	AUC	Type II Error	F-score
ROS	RUS	0.771	0.276	0.460	**0.782**	0.293	0.507
	Tomek-Links	0.771	0.531	0.434	0.778	0.541	0.477
	ENN	**0.773**	0.242	**0.465**	0.779	0.286	0.500
	OSS	0.771	0.535	0.431	0.778	0.534	0.481
	NCL	0.771	0.335	0.459	0.776	0.331	**0.520**
	NearMiss-1	0.717	0.363	0.411	0.727	0.331	0.473
	NearMiss-2	0.731	0.331	0.420	0.749	0.316	0.473
	NearMiss-3	0.763	**0.106**	0.415	0.773	**0.113**	0.436
SMOTE	RUS	0.760	0.270	0.456	0.767	0.286	0.488
	Tomek-Links	0.761	0.538	0.432	0.762	0.579	0.448
	ENN	0.766	0.244	0.464	0.776	0.256	0.495
	OSS	0.760	0.538	0.433	0.763	0.579	0.448
	NCL	0.762	0.307	0.454	0.764	0.338	0.484
	NearMiss-1	0.746	0.322	0.440	0.750	0.293	0.491
	NearMiss-2	0.739	0.311	0.425	0.745	0.353	0.435
	NearMiss-3	0.736	0.171	0.412	0.745	0.135	0.451
ADASYN	RUS	0.753	0.279	0.443	0.767	0.241	0.506
	Tomek-Links	0.753	0.546	0.418	0.760	0.571	0.445
	ENN	0.764	0.251	0.452	0.776	0.233	0.498
	OSS	0.753	0.544	0.421	0.760	0.571	0.447
	NCL	0.756	0.296	0.452	0.763	0.271	0.500
	NearMiss-1	0.738	0.309	0.434	0.761	0.271	0.489
	NearMiss-2	0.717	0.296	0.416	0.736	0.286	0.467
	NearMiss-3	0.724	0.169	0.400	0.730	**0.105**	0.426
Borderline-SMOTE	RUS	0.759	0.270	0.456	0.773	0.248	0.510
	Tomek-Links	0.759	0.514	0.441	0.767	0.579	0.448
	ENN	0.768	0.251	0.459	0.776	0.211	0.510
	OSS	0.759	0.520	0.438	0.767	0.586	0.442
	NCL	0.761	0.302	0.454	0.767	0.293	0.491
	NearMiss-1	0.749	0.311	0.438	0.761	0.271	0.497
	NearMiss-2	0.728	0.315	0.417	0.737	0.293	0.457
	NearMiss-3	0.734	0.164	0.415	0.742	0.113	0.456

the performance in terms of F-score and AUC, they also result in a high increase with regard to Type II error. Therefore, the Python packages' previously mentioned default values are used to fit the final model.

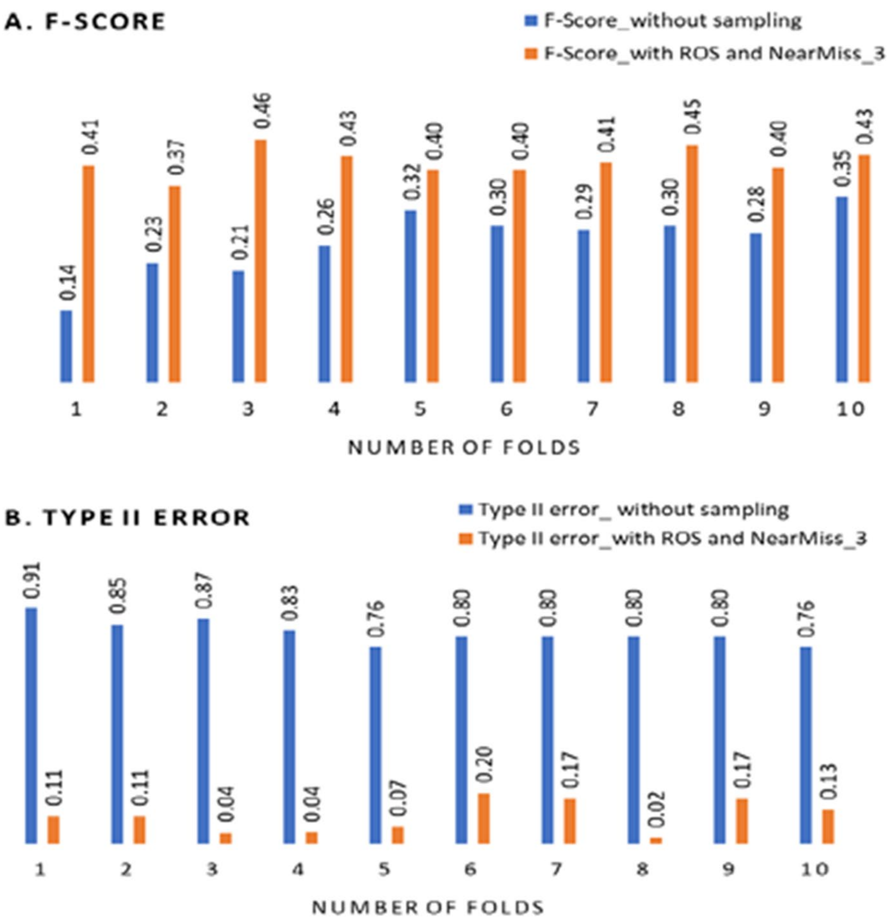

Fig. 2 Cross-validation for (**A**) F-scores and (**B**) Type II error of the Logistic classifier before sampling and after applying the hybrid of ROS and NearMiss-3 – created by researchers

Table 7 Hyper-parameter tuning setting values

Technique	Parameters	Best Values
LR	Penalty = [‘l1’, ‘l2’, None]	Penalty = l2
ROS	sampling_strategy (targeted balance ratio) = [0.3, 0.5, 0.7]	sampling_strategy = 0.5
NearMiss-3	sampling_strategy = [0.8, 0.9, 1.0] n_neighbors (k) = [1, 3, 5, 7, 9]	sampling_strategy = 1.0 k = 7

5.5 Final logistic classification model

Figure 3 presents the learning curves of the Logistic classifier for the recall (i.e., (1 – Type II error)) and F-score metrics after applying the combination of *ROS* and

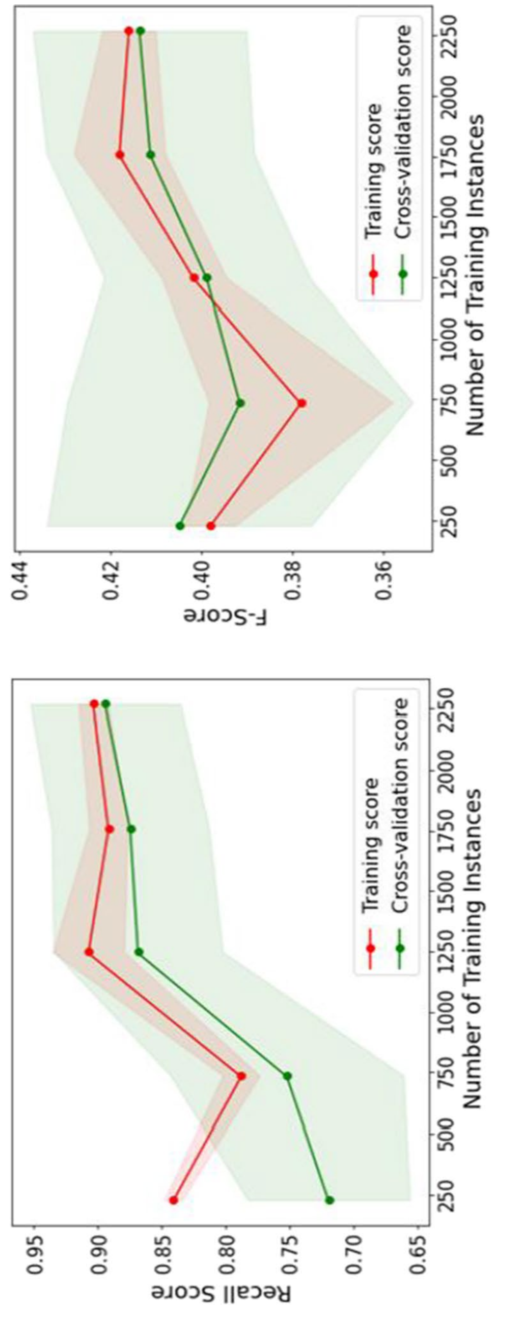

Fig. 3 Learning curves for recall and F-score for Logistic classifier with ROS and NearMiss-3 – created by researchers

Table 8 Final Logistic model of school-dropout

Variables	Coef.β_j	Odds = $exp(\beta_j)$	Variables	Coef.β_j	Odds = $exp(\beta_j)$
Intercept	- 0.8129699		Equal_treatment	- 0.2705161	0.7629856
Parents_illitracy	0.5657427	1.7607549	Class_fail	0.5160966	1.6754748
Poverty	0.2803968	1.3236549	Year_ repetition	- 0.2163406	0.8054609
Chronic diseases	0.7291549	2.0733276	Teacher_caring	- 0.3779855	0.6852405
Nursery	- 0.2188644	0.8034306	Siblings_2	0.1193805	1.1267986
Punishment	0.2705889	1.3107361	Siblings_3	0.2452311	1.2779166
Co_educational	0.7164258	2.0471034	Birth_order_2	- 0.0071561	0.9928695
Shifts	- 0.1817539	0.8338065	Birth_order_3	0.0992124	1.1043008

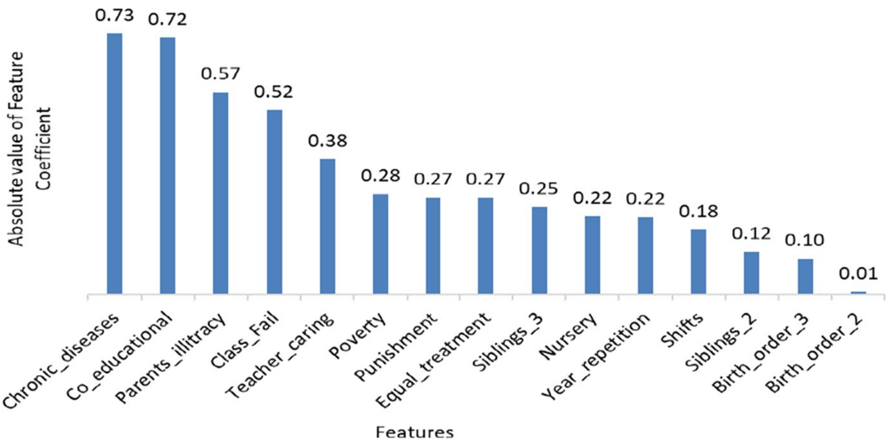

Fig. 4 Importance of respondent features on school-dropouts based on $|\beta_j|$

NearMiss-3. It shows that the training and cross-validation scores converge together as more data is added, therefore the model will probably not benefit from adding more training datasets. In other words, investigating more data will not improve the results. Consequently, the final model is fitted on the entire dataset after being re-sampled through *ROS* and *NearMiss-3*.

Table 8 shows the final model's coefficients that fit the Logistic regression equation. The Chi-Square (χ^2) test of the null hypothesis stating that all coefficients on this fitted model are equal to zero is rejected at 95% confidence.

When all other variables are held constant, the general interpretation of the odds of being at-risk of any category of a variable x_j, is the reported value in Table 8 times greater than that of being at-risk in the reference category of that variable (Hosmer et al., 2013). For the binary variables, the category coded 0 is considered the reference category.

Based on the absolute values of the Logistic model coefficients $|\beta_j|$, Fig. 4 reports the explanatory features according to their importance. It shows that student chronic diseases, co-educational, parents' illiteracy, educational performance, and teacher caring represent the most important five features affecting the school-dropout problem in Egypt.

6 Discussions of the main results

As important as it is for policymakers to comprehend the reasons why students drop out, they must first determine who is typically at-risk. This is essential so as to come up with practical interventions where they are mostly needed, especially in environments with limited resources and competing objectives, which characterize the majority of the world's educational systems (Moreno & Hector, 2018). However, early identifying at-risk students is not an easy task because it is challenging to characterize the at-hand problem by a few defining features. Consequently, as detailed earlier in the introduction, this study seeks to develop a Logistic model having an ability to early identify students at-risk of school-dropout in Egypt.

Following the data manipulation, a descriptive analysis is conducted to quickly examine the key characteristics of the surveyed respondents. As shown in Table 9, out of the total sample of 3154, 19% are dropout cases. Among them, 69% are female students and 63% are living in rural areas. However, consistently with what figures out by Mali et al. (2012), neither gender nor place of residence is significantly correlated with school-dropout (see Table 4).

Overall, based on the coeffecients' values presented in Table 8, the decision rule to decide that a student is likely to dropout is $p\left(y_i = 1\middle|X_i\right) = \frac{exp(W)}{1+exp(W)}$, where:

$$\begin{aligned}
W = {} & -0.81 + 0.57 * (Parents_illitracy) \\
& + 0.28 * (Poverty) + 0.73 * (Chronic_disease) \\
& - 0.22 * (Nursery) \\
& + 0.27 * (Punishment) + 0.72 * (Co_educational) \\
& - 0.81 * (Shifts) - 0.27 * (Equal_treatment) \\
& + 0.52 * (Class_fail) - 0.23 * (Year_repetition) \\
& - 0.38 * (Teacher_caring) + 0.12 * (Siblings_2) \\
& + 0.25 * (Siblings_3) - 0.01 * (Birth_order_2) \\
& + 0.10 * (Birth_order_3).
\end{aligned}$$

Keeping in mind that, if $p\left(y_i = 1|X_i\right) > 0.5$, the enrolled student is, by default, assumed likely to dropout. However, the decision-maker could choose a lower probability threshold. Lowering the threshold resulted in increasing the number of correctly predicted dropouts, meanwhile the overall model accuracy decreases because many of those who are really non-dropouts are misclassified as dropouts. In this situation, the educational authorities may encounter significant financial costs as a result of intervening with students who are not truly at-risk. Berens et al. (2019) suggests assigning the threshold based on the typical dropout rate of students who were enrolled in previous years.

The above fitted Logistic model (see also Fig. 4) indicates that student chronic diseases, co-educational, parents' illiteracy, educational performance, and teacher caring are manifested as the top five determinants imposing vulnerability to dropout among the Egyptian students in the compulsory level. The majority of these variables are matching with those figured out by other studies in Egypt and developing countries as well like India and Tanzania (Elbadawy, 2014; Mnyawami et al., 2022; Rahaman & Das, 2018). It is worth mentioning that these two countries are classified by the World Bank as lower middle-income developing countries, and their Human Capital Index scores in 2020 are 0.5 and 0.4 respectively compared to 0.5 for Egypt.

Taken as a whole, explanations and implications of these results could be summarized as follows. First, the Logistic model reveals that chronic diseases represent the most important feature affecting school-dropout. In this context, it is found that having a chronic disease nearly doubles the odds of being an at-risk student, when compared to students with no chronic diseases. Tate (2013) clarifies that because there is a link between health and education, young people's health affects their ability to learn/study or to succeed in school, particularly in the case of illnesses that begin in childhood and last a lifetime. Chronic diseases usually limit students' ability to complete school in most cases. To address this problem, policymakers may employ several health-related interventions to raise education attainment and to reduce dropout rates. Coordinated school health programmes, health clinics, and mental health services are a few examples of these school-based initiatives.

Second, when compared to mixed (co-educational) schools, single-sex schools typically have a lower dropout rate. This could be a result of the societal norms and traditions that persist in many developing countries and that act as a barrier to the education of girls. Girls and boys are generally regarded differently in developing countries due to perceptions and expectations about their roles in the household, the workplace, and the wider society. This, in turn, has an impact on how families choose to educate their children. Similar results are reported by Badr (2012).

Third, when both parents are illiterate, it is found that the odds of dropping out of school increases by roughly 1.8 times more than when at least one of them is not. This is because students of uneducated parents are left without support in their schooling. Similar findings are also reported by Elbadawy (2014).

Fourth, the model supports the findings of Weybright et al. (2017) by pointing that educational success has a considerable impact on the likelihood of dropping out. Dropout rates are higher for students who struggle in the classroom and lagging behind. Failing a class increases the odds that a student will drop out by approximately 1.7. In reality, early intervention with those youngsters may help them stay in school and avoid repetitive failure. It could be beneficial for educational institutions to have academic advisors check in with the students periodically throughout the educational year. Students in the initial stages of their education typically need a mentor to provide them with educational and emotional support on occasion.

Fifth, teacher caring is proved to have a direct influence on the problem of concern. It follows that students who have teachers providing them with counsel and showing concern for their struggles are less likely to drop out than those who do not.

This finding should encourage policymakers and school administrators to promote constructive communication between students and teachers. This can be accomplished by designating such units and committees in charge of carrying out this task so as to prompt communication between students and their teachers to inspire students towards succuss.

Goudet et al. (2017) clarify that poverty places a heavy burden on the family, and malnutrition causes major obstacles in poor children's physical and mental development, which makes it difficult for them to keep up with the demands of school and forces them to drop out. Also, Timbal (2019) figures out that number of student's living siblings is an important factor affecting the risk of dropping out. However, both variables have no similar importance in the fitted Logistic model for the Egyptian case. Poverty comes sixth in order, while siblings_3 comes ninth.

7 Conclusions

Even though the Egyptian government investments in the education sector has increased in order to make compulsory basic education universal, the persistent phenomenon of school-dropout continues to be one of the challenging problems. In general, the majority of school-dropouts typically experience chronically higher unemployment rates owing to the fact that they suffer from lack of knowledge, skills, and intellectual capabilities necessary to compete in today's job market and get accepted in contemporary society. Dropouts are often less motivated educationally, and commonly have complex psychological and behavioural issues that put them in danger. Consequently, the dropout issue is upsetting and poses a threat to successfully completing schooling as well as to attaining the overall objectives of education. As a result, it is crucial to determine the causes of dropout to provide policymakers with guidance so as to eradicate this social behavior over time and to address the issue of unifying basic education that is required of all students (Gubbels et al., 2019; Rahaman & Das, 2018).

In this perspective, the current research study attempts to address the actual dropout problem in Egypt by developing a Logistic model having an ability to early predict students at-risk of dropping out at basic education, while dealing with the class-imbalance problem to improve the model's overall performance. To accomplish this goal, the study investigates the imbalance issue through a comparative analysis of common under- and over-sampling techniques, as well as their combinations in order to decrease Type II error whereas maintaining both AUC and F-score at acceptable levels.

Resampling is critical to be considered when there is a class-imbalance problem in the dataset under investigation. Nevertheless, the performance of the techniques used for resampling are highly affected by the dataset's characteristics. In some cases, to improve the classifier performance, a combination of both under- and over-sampling techniques is required. For the school-dropout problem, as evidenced by the study's findings, when a combination of under- and over-sampling techniques is employed, the performance scores of the Logistic classifier outperforms individual sampling techniques and the model with the dataset without resampling, especially in terms of F-score and Type II error. Being more specific, the results show that the *ROS* combined with *NearMiss-3* is thought to be improving the performance when integrated with the Logistic classifier as compared with other resampling strategies. This is because it produces the lowest Type II and also results in high AUC and F-score. However, these results are limited to the dataset of the specific subject of concern.

As for the model's essential results, it is shown that student chronic diseases, co-educational, parents' illiteracy, educational performance, and teacher caring are the main five factors making Egyptian students enrolled in the compulsory level vulnerable to dropping out. As a result, prompt action and attention from policymakers or school administrators to these features may help to resolve the issue early on. This could assist in preventing many students from quitting school which represents, as mentioned before, a severe problem that limits their involvement in economic and societal activities for the rest of their lives.

One of the current study's shortcomings is that the dataset is relatively old, however as far as the researchers know, it is the most recent national survey covering the issue of school-dropout. Finally, for further improving the classification performance of school dropouts, there are some additional options that could be investigated in future research, such as more hyper-tuning the model parameters and implementing other classification techniques such as Decision Trees and Support Vector Machine. Another potential research direction hindering on the availability of data is to consider other important behavioural and psychological attributes.

Appendix 1

Table 9 Description of the selected variables

Variable Name and Description	Feature Labels	Dropouts 19% (N = 596)	Non-Dropouts 81% (N = 2558)
Dropout: This is the target variable which indicates whether or not the instance is a school dropout	0: non-dropout 1: dropout		
Gender	0: female	69%	70%
	1: male	31%	30%
Place of residence	0: rural	63%	61%
	1: urban	37%	39%
Parents illiteracy	0: at least one of individual's parents is educated	28%	57%
	1: both are illiterate	72%	43%
Father's work status	0: not work	6%	6%
	1: work	94%	94%
Mother's work status	0: not work	91%	91%
	1: work	9%	9%
Poverty: Respondents are divided into two groups based on the Wealth Index Quintiles. The first was referred to as "poor", and it included both the poorest and poorer categories. The "non-poor" includes the remaining categories	0: non-poor	51%	75%
	1: poor	49%	25%
Siblings: The number of individual's living brothers/sisters	1: less than or equal two (reference category)	9%	15%
	2: greater than or equal three and less than five	33%	48%
	3: greater than or equal five	58%	37%

Table 9 (continued)

Variable Name and Description	Feature Labels	Dropouts 19% (N = 596)	Non-Dropouts 81% (N = 2558)
Birth order: The individual's birth order among all living siblings	1: less than or equal two (reference category)	44%	53%
	2: greater than or equal three and less than five	33%	31%
	3: greater than or equal five	23%	16%
Chronic diseases	0: not have any chronic diseases	93%	96%
	1: have any chronic diseases	7%	4%
Nursery: Whether or not the individual was enrolled in a pre-primary stage nursery	0: no	92%	78%
	1: yes	8%	22%
Punishment: Whether or not students at individual's last basic-level school were subjected to physical punishment	0: no	23%	40%
	1: yes	77%	60%
Co-educational: Whether or not unisex enrollment was permitted at the individual's last basic-level school	0: not unisex	16%	46%
	1: unisex	84%	54%
Shifts: Whether or not individual's last basic-level school was operated in shifts	0: single period	65%	70%
	1: operated in shifts	35%	30%

Table 9 (continued)

Variable Name and Description	Feature Labels	Dropouts 19% (N = 596)	Non-Dropouts 81% (N = 2558)
Equal treatment: The rules at school are equally applied to all students without discrimination	0: not equally applied	81%	76%
	1: equally applied	19%	24%
Private tutoring: Whether or not the individual utilized any private tutoring at the last basic-level school	0: not utilize	0%	0%
	1: utilize	100%	100%
Class-fail: Whether or not the individual failed a class and repeated an examination in any year	0: not fail	75%	88%
	1: fail	25%	12%
Year repetition: Whether or not the individual repeated any school year	0: not repeat	83%	91%
	1: repeat	17%	9%
Teacher caring: Teachers care about the students' personal problems and help resolve them	0: not care	92%	85%
	1: care	8%	15%

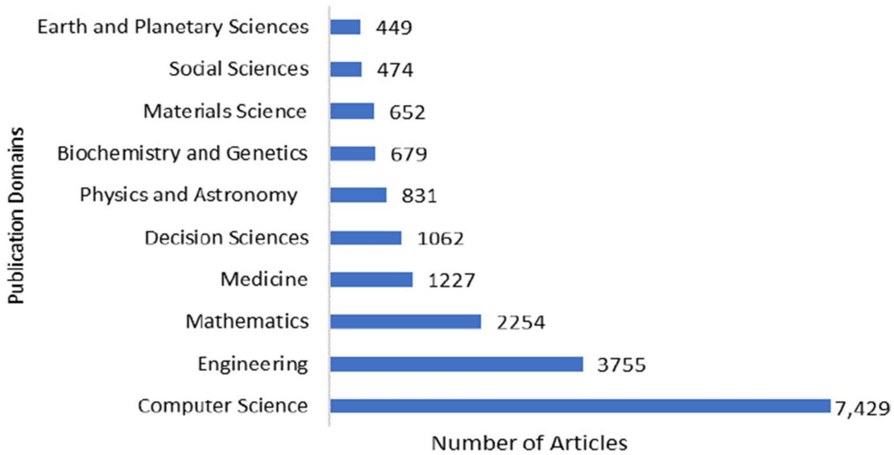

Fig. 5 Top 10 domains in which most imbalanced learning papers were published

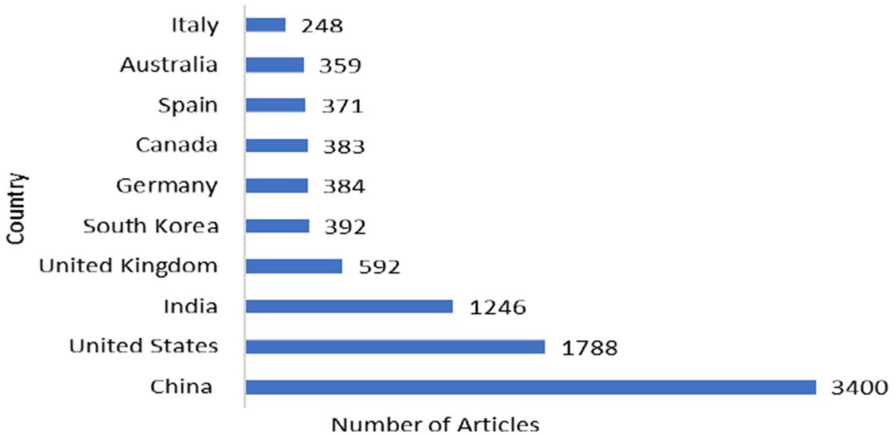

Fig. 6 Top 10 countries of publications in imbalance learning

Acknowledgements We wish to express our sincere gratitude to the anonymous reviewers for taking the time and effort necessary to review the manuscript. The reviewers' valuable comments and recommendations significantly have assisted the authors to improve the manuscript's quality.

Funding Open access funding provided by The Science, Technology & Innovation Funding Authority (STDF) in cooperation with The Egyptian Knowledge Bank (EKB).

Data availability The original dataset of this study is available upon request from Harvard Dataverse through the following link. https://doi.org/10.7910/DVN/89Y8YC

Declarations

Conflict of interest The authors declare no relevant financial or non-financial competing interests.

References

Agustianto, K., & Destarianto, P. (2019). Imbalance Data Handling using Neighborhood Cleaning Rule (NCL) Sampling Method for Precision Student Modeling. *International Conference on Computer Science, Information Technology, and Electrical Engineering, ICOMITEE*, 86–89.

Amin, A., Anwar, S., Adnan, A., Nawaz, M., Howard, N., Qadir, J., Hawalah, A., & Hussain, A. (2016). Comparing Oversampling Techniques to Handle the Class Imbalance Problem: A Customer Churn Prediction Case Study. *IEEE Access, 4*, 7940–7957.

Assaad, R. (2010). The Effect of Domestic Work on Girls' Schooling: Evidence from Egypt. *Feminist Economics, 16*(1), 79–128.

Avon, V. (2016). *Machine learning techniques for customer churn prediction in banking environments*. University of Padua. An M.Sc. thesis retrieved from https://core.ac.uk/download/pdf/83461632.pdf. Accessed 12 June 2021.

Badr, M. (2012). School Effects on Educational Attainment in Egypt. *CREDIT Research Paper, 12*(5), 1–58.

Berens, J., Schneider, K., Görtz, S., Oster, S., & Burghoff, J. (2019). Early Detection of Students at Risk – Predicting Student Dropouts Using Administrative Student Data and Machine Learning Methods. *Journal of Educational Data Mining, 11*(3), 1–41.

Berrar, D. (2018). Bayes' Theorem and Naive Bayes Classifier Bayes. In *Encyclopedia of Bioinformatics and Computational Biology* (pp. 403–412). Elsevier Science Publisher.

Chau, V. T. N., & Phung, N. H. (2013). Imbalanced Educational Data Classification: An Effective Approach with Resampling and Random Forest. *International Conference on Computing and Communication Technologies: Research, Innovation, and Vision for Future, RIVF*, 135–140.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique Nitesh. *Journal of Artificial Intelligence Research, 16*, 321–357.

Elbadawy, A. (2014). *Education in Egypt: Improvements in Attaiment Problems with Quality and Inequality* (Economic Research Forum (ERF) Working Paper 854).

Elreedy, D., & Atiya, A. F. (2019). A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance. *Information Sciences, 505*, 32–64.

Ghorbani, R., & Ghousi, R. (2020). Comparing Different Resampling Methods in Predicting Students ' Performance Using Machine Learning Techniques. *IEEE Access, 8*, 67899–67911.
Goel, G., Maguire, L., Li, Y., & McLoone, S. (2013). Evaluation of Sampling Methods for Learning from Imbalanced Data. *International Conference on Intelligent Computing*, 392–401.
Gonzalez-Abril, L., Angulo, C., Nuñez, H., & Leal, Y. (2017). Handling Binary Classification Problems with a Priority Class by Using Support Vector Machines. *Applied Soft Computing Journal, 61*, 661–669.
Goudet, S. M., Kimani-Murage, E. W., Wekesah, F., Wanjohi, M., Griffiths, P. L., Bogin, B., & Madise, N. J. (2017). How does poverty affect children's nutritional status in Nairobi slums? A qualitative study of the root causes of undernutrition. *Public Health Nutrition, 20*(4), 608–619.
Gubbels, J., van der Put, C. E., & Assink, M. (2019). Risk Factors for School Absenteeism and Dropout: A Meta-Analytic Review. *Journal of Youth and Adolescence, 48*(9), 1637–1667.
Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from Class-Imbalanced Data: Review of Methods and Applications. *Expert Systems with Applications, 73*, 220–239.
Hanushek, E. A., Lavy, V., & Kohtaro, H. (2006). Do Students Care about School Quality? Determinants of Dropout Behavior in Developing Countries. In *NBER Working Paper* (Issue 12737).
Hasan, M. N. (2019). A Comparison of Logistic Regression and Linear Discriminant Analysis in Predicting of Female Students Attrition from School in Bangladesh. *4th International Conference on Electrical Information and Communication Technology (EICT)*, 1–3.
He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. *IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 1322–1328.
He, H., & Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263–1284.
Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression*. Wiley & Sons Inc.
Hsu, J. L., Hung, P. C., Lin, H. Y., & Hsieh, C. H. (2015). Applying Under-Sampling Techniques and Cost-Sensitive Learning Methods on Risk Assessment of Breast Cancer. *Journal of Medical Systems, 39*(4), 1–13.
Kabathova, J., & Drlik, M. (2021). Towards Predicting Student's Dropout in University Courses Using Different Machine Learning Techniques. *Applied Sciences, 11*(1), 1–19.
Koutina, M., & Kermanidis, K. L. (2011). Predicting Postgraduate Students' Performance Using Machine Learning Techniques. *International Conference on Engineering Applications of Neural Networks*, 159–168.
Kraiem, M. S., Sánchez-Hernández, F., & Moreno-García, M. N. (2021). Selecting the Suitable Resampling Strategy for Imbalanced Data Classification Regarding Dataset Properties. An Approach Based on Association Models. *Applied Sciences, 11*(18), 1–26.
Kristoffersen, L. R., & Hernandez, R. M. (2021). A Comparative Performance of Breast Cancer Classification Using Hyper-Parameterized Machine Learning Models. *International Journal of Advanced Technology and Engineering Exploration, 8*(82), 1080–1101.
Kubat, M., & Matwin, S. (1997). Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. *International Conference on Machine Learning, 97*, 179–186.
Laurikkala, J. (2001). Improving Identification of Difficult Small Classes by Balancing Blass Distribution. *Conference on Artificial Intelligence in Medicine in Europe*, 63–66.
Liang, D., Tsai, C. F., Dai, A. J., & Eberle, W. (2018). A Novel Classifier Ensemble Approach for Financial Distress Prediction. *Knowledge and Information Systems, 54*(2), 437–462.
Lloyd, C. B., Tawila, S. El, Clark, W. H., & Mensch, B. (2001). Determinants of Educational Attainment Among Adolescents in Egypt : Does School Quality Make a Difference ? In *Policy Research Division Working Paper* (Issue 150).
Loyola-González, O., Martínez-Trinidad, J. F., Carrasco-Ochoa, J. A., & García-Borroto, M. (2016). Study of the Impact of Resampling Methods for Contrast Pattern Based Classifiers in Imbalanced Databases. *Neurocomputing, 175*, 935–947.
Maimon, O., & Rokach, L. (2015). *Data Mining with Decision Trees: Theory and Applications*. World Scientific Publishing Co.
Mali, S., Patil, D. M., & Manaspure, S. P. (2012). A comparative Study of The School Dropouts with a Socio-Demographically Comparison Group of Urban Slum Inhabitants in Maharashtra. *International Journal of Biomedical and Advance Research, 3*(5), 329–335.

Mani, I., & Zhang, I. (2003). KNN Approach to Unbalanced Data Distributions: A Case Study Involving Information Extraction. *Proceedings of Workshop on Learning from Imbalanced Datasets, International Conference on Machine Learning (ICML)*, *126*, 1–7.

Mduma, N., Kalegele, K., & Machuve, D. (2019). Machine Learning Approach for Reducing Students Dropout Rates. *International Journal of Advanced Computer Research, 9*(42), 156–169.

Mnyawami, Y. N., Maziku, H. H., & Mushi, J. C. (2022). Enhanced Model for Predicting Student Dropouts in Developing Countries Using Automated Machine Learning Approach: A Case of Tanzanian's Secondary Schools. *Applied Artificial Intelligence, 36*(1), 432–451.

Mohammed, A. J. (2020). Improving Classification Performance for a Novel Imbalanced Medical Dataset using SMOTE Method. *International Journal of Advanced Trends in Computer Science and Engineering, 9*(3), 3161–3172.

Mohammed, R., Rawashdeh, J., & Abdullah, M. (2020). Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results. *11th International Conference on Information and Communication Systems, ICICS 2020*, *May*, 243–248.

Moreno, M., & Hector, A. (2018). Predicting School Dropout with Administrative Data New Evidence from Guatemala and Honduras. *Education Economics, 26*(4), 356–372.

Napierala, K., & Stefanowski, J. (2012). BRACID: A Comprehensive Approach to Learning Rules from Imbalanced Data. *Journal of Intelligent Information Systems, 39*(2), 335–373.

Nguyen, H. M., Cooper, E. W., & Kamei, K. (2011). Borderline Over-Sampling for Imbalanced Data Classification. *International Journal of Knowledge Engineering and Soft Data Paradigms, 3*(1), 4–21.

Orooji, M., & Chen, J. (2019). Predicting Louisiana Public High School Dropout through Imbalanced Learning Techniques. *18th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 456–461.

Peng, C.-Y.J., So, T.-S.H., Stage, F. K., John, E. P., & St. (2002). The Use and Interpretation of Logistic Regression in Higher Education Journals: 1988–1999. *Research in Higher Education, 43*(3), 259–293.

Population Council. (2015). *Survey of Young People in Egypt (SYPE) 2014*. Retrieved from: https://www.unicef.org/egypt/media/4976/file/2014_Survey_on_Young_People_in_Egypt.pdf. Accessed 20 June 2022

Quadri, M. N., & Kalyankar, N. V. (2010). Drop Out Feature of Student Data for Academic Performance Using Decision Tree Techniques. *Global Journal of Computer Science and Technology, 10*(2), 2–5.

Radwan, A., & Cataltepe, Z. (2017). Improving Performance Prediction on Education Data with Noise and Class Imbalance. *Intelligent Automation & Soft Computing, 8587*, 1–8.

Radwan, M. (2019). Causes of the Phenomenon of School Dropout among Girls and its Impacts in Rural Areas of EL-Ayat District, Giza Governorate, Egypt. *Egyptian Journal of Agricultural Sciences, 70*(2), 91–101.

Rahaman, M., & Das, D. N. (2018). Determinants of School Dropouts in Elementary Education in Manipur. *Indian Journal of Geography and Environment, 15*(16), 89–106.

Rashu, R. I., Haq, N., & Rahman, R. M. (2014). Data Mining Approaches to Predict Final Grade by Overcoming Class Imbalance Problem. *17th International Conference on Computer and Information Technology, ICCIT*, 14–19.

Ratih, I. D., Retnaningsih, S. M., Islahulhaq, I., & Dewi, V. M. (2022). Synthetic Minority Over-Sampling Technique Nominal Continous Logistic Regression for Imbalanced Data. *American Institute of Physics (AIP) Conference Proceedings*, *2668*(1).

Safaa, E., & El-Daw, A. S. (2001). Poverty, human capital and gender: A comparative study of Yemen and Egypt. In *Economic Research Forum Working Paper* (Issue 0123). https://erf.org.eg/publications/poverty-human-capital-gender-comparative-study-yemen-egypt/. Accessed 23 Nov 2021.

Sarra, A., Fontanella, L., & Di Zio, S. (2019). Identifying Students at Risk of Academic Failure Within the Educational Data Mining Framework. *Social Indicators Research, 146*(1), 41–60.

Shamsudin, H., Yusof, U. K., Jayalakshmi, A., & Akmal Khalid, M. N. (2020). Combining Oversampling and Undersampling Techniques for Imbalanced Classification: A Comparative Study Using Credit Card Fraudulent Transaction Dataset. *IEEE International Conference on Control and Automation, ICCA*, 803–808.

Suliman, E. D. A., & El-kogali, S. E. (2002). Why Are the Children out of School?: Factors Affecting Children's Education in Egypt. *Ninth Economic Research Forum (ERF) Annual Conference*, 26–28.

Tansey, R., White, M., Long, R. G., & Smith, M. (1996). A Comparison of Loglinear Modeling and Logistic Regression in Management Research. *Journal of Management, 22*(2), 339–358.

Tate, W. F. (2013). How Does Health Influence School Dropout? In *A report on the health and well-being of African Americans in St. Louis*. Washington University.
Thai-Nghe, N., Busche, A., & Schmidt-Thieme, L. (2009). Improving Academic Performance Prediction by Dealing with Class Imbalance. *9th International Conference on Intelligent Systems Design and Applications*, 878–883.
Timbal, M. A. (2019). Analysis of Student-at-Risk of Dropping out (SARDO) Using Decision Tree: An Intelligent Predictive Model for Reduction. *International Journal of Machine Learning and Computing, 9*(3), 273–278.
Tomek, I. (1976). Two Modifications of CNN. *IEEE Transactions on Systems, Man, and Cybernetics, 6*, 769–772.
UNICEF. (2017). *Early Warning Systems for Students at Risk of Dropping out* (UNICEF Series on Education Participation and Dropout Prevention).
Weybright, E. H., Caldwell, L. L., Wegner, L., & Smith, E. A. (2017). Predicting secondary school dropout among South African adolescents: A survival analysis approach. *South African Journal of Education, 37*(2), 1–11.
Wilson, D. L. (1972). Asymptotic Properties of Nearest Neighbor Rules Using Edited Data. *IEEE Transactions on Systems, Man and Cybernetics, 2*(3), 408–421.
Yehuala, M. A. (2015). Application of Data Mining Techniques for Student Success and Failure Prediction (The Case Of Debre_Markos University). *International Journal of Scientific & Technology Research, 4*(4), 91–94.
Yi, X., Xu, Y., Hu, Q., Krishnamoorthy, S., Li, W., & Tang, Z. (2022). ASN-SMOTE: A Synthetic Minority Oversampling Method with Adaptive Qualified Synthesizer Selection. *Complex & Intelligent Systems*. https://doi.org/10.1007/s40747-021-00638-w