Towards a Students' Dropout Prediction Model in Higher Education Institutions Using Machine Learning Algorithms

https://doi.org/10.3991/ijet.v17i18.25567

Khalid Oqaidi[1(✉)], Sarah Aouhassi[2], Khalifa Mansouri[1]
[1]Laboratory SSDIA, ENSET of Mohammedia, University Hassan II of Casablanca, Mohammedia, Morocco
[2]Laboratory SSDIA, ENSAD, University Hassan II of Casablanca, Mohammedia, Morocco
khalid.oqaidi@gmail.com

Abstract—Using machine learning to predict students' dropout in higher education institutions and programs has proven to be effective in many use cases. In an approach based on machine learning algorithms to detect students at risk of dropout, there are three main factors: the choice of features likely to influence a partial or total stop of the student, the choice of the algorithm to implement a prediction model, and the choice of the evaluation metrics to monitor and assess the credibility of the results. This paper aims to provide a diagnosis of machine learning techniques used to detect students' dropout in higher education programs, a critical analysis of the limitations of the models proposed in the literature, as well as the major contribution of this article is to present recommendations that may resolve the lack of global model that can be generalized in all the higher education institutions at least in the same country or in the same university.

Keywords—student retention, dropout prediction, higher education, machine learning, classification

1 Introduction

Through a bibliographical study we underline the essential elements in students' dropout prediction study. Then we highlight some issues in the literature such as:

- The lack of a global model that brings together all the possible variables that concern the students subject of the study;
- The lack of an ease of choice of one machine learning algorithm over another according to the objective of the study: in fact, the algorithms used in different use cases differ according to the variables and the data available;
- Lack of clear criteria of the choice of the evaluation metric to evaluate the machine learning algorithm chosen.

This article is organized in sections. In section 2 we present the literature review and comment the results extracted. In section 3 we discuss and highlight the missing points

in the literature and we present our recommendations to complement them in section 4. We conclude with section 5.

In our methodology we selected 29 recent articles, most of them dated between 2018 and 2020. In order to be located within the subject, we started with 2 articles on bibliographic studies of works on the dropout and performance prediction of students in higher education institutions. Then 1 article on the multi-criteria aspect of the choice of the algorithm, input variables and metrics for evaluating machine learning models in different contexts. The following 25 articles are use cases of implementing machine learning models to predict dropout in the context of higher education, presented in chronological order of publication. The 29th article is a reference of the confusion matrix that we present in Table 2. The choice of the articles on dropout and performance is based on the fact that all the articles must focused on higher education programs which are degree oriented. This has been achieved by avoiding all studies that are interested in business-oriented online courses that are not geared towards issuing a university degree, and all studies on primary or secondary education. The articles were found using the key words: students' dropout, higher education, machine learning, prediction and students' performance in databases as Elsevier, IEEE Xplore, ACM, Springer and other journals.

2 Literature review

2.1 Related works

In this survey the authors analyze different contributions of students' dropout prediction in India between 2009 and 2016, and try to localize the missing elements that make the gaps between the previous studies. They stressed four kinds of studies in Educational Data Mining: Classification, Clustering, Prediction and Association Rule mining. The machine learning classifiers found in the literature are varied, we note the most used: Support Vector Machine, Decision Tree algorithms, Artificial Neural Networks, Logistic Regression, Naïve Bayes, Random Forest and others. The data variables used to implement the models are diversified, we note some of them: grade in high school, secondary school and other related education, Gender, Family structure, Parents Qualification, Parents Occupation, Required for Household work, Addictions (Alcohol, Smoke, Pills, Solvents, Drugs etc.), Basic facility in the education institution different for boys and girls, Poor Teaching methodology adopted, Got married [1].

1681 identified papers 67 selected ones to write a systematic literature according to the Organization for Economic Cooperation and Development (OECD), the European dropout rates ranged between 30% and 50%, while in the USA the student dropout rate was 37%. They identified the techniques used for data pre-processing, the factors affecting dropout, the techniques used for factor selection, the techniques used for prediction, their levels of reliability and the tools used [2].

The selection and evaluation of the machine learning algorithm between the large choices of possibilities consume lot of time when it is done manually, and usually give not the results wanted. A classifier must be chosen regarding not only the accuracy, but also time complexity and consistency [3].

Based on the assumption that a single algorithm may fail to detect accurately the dropout of e-learning courses, the authors have used three machine learning algorithms for the same purpose. The three methods have been combined and evaluated through accuracy, sensitivity and precision. The results gave more satisfaction than each of the algorithms alone [4].

Models that classify students at risk of abandoning a degree were implemented on 27 university degrees using logistic regression. The goal was to locate the variables responsible of dropout, the most important among them are: start age, parents' studies, academic performance, success, average mark in the degree and others. The authors have noticed that at the same university, two different degrees may have different reasons of dropout [5].

A comparative study to know the optimal Decision Tree to predict students' performance was made. This study is on only just one machine learning algorithm and its goal was the comparison between these variants of decision tree using metrics such as accuracy and computational time. The results state that J48 tree was the most suitable algorithm [6].

A predictive model in e-learning context was developed using three machine learning algorithms: Artificial Neural Network, Decision Tree and Bayesian Networks, with two categories of features: students' personal characteristics and academic performance. The metrics used to evaluate the models are accuracy, precision, recall and F-measure. Decision tree gave the best results in this case and the students at risk of dropout have been addressed to the department of the university concerned to take the necessary measures [7].

A personalized course recommendations based on machines learning techniques oriented system was proposed. The system makes the prediction of students' dropout as well as the of the students course grades. The data used has the particularity of containing only the subject marks obtained in the current year and the entry grade score to the current institution [8].

An Early Detection System (EDS) was developed in order to detect students at risk of dropping out by predicting their performance in an early stage. The algorithms Regression Analysis, Neural Networks, Decision Trees and the AdaBoost are used to identify the factors responsible for a probable dropout in German universities. Two categories of data were extracted to train the model: demographic, historical performance and current academic performance. The accuracy was higher at the end of the fourth semester than the first semester, and the AdaBoost performed better than the other algorithms [9].

Two machine learning algorithms have been implemented, Decision Tree and random forest, to predict students' dropout between first and second years in an institute of technology. The model focused on academic and institutional features only [10].

The authors collected data via questions survey on different factor classes like Academics, Demographical factors, Psychological factors, Health issues and Behavioral factors. They implemented the model using Naïve Bayes classifier as only algorithm and evaluated it with metrics such as Recall and Precision [11].

Unlike most studies, the authors considered 37 non-academic factors for the dropout prediction. These factors are grouped in 5 categories: demographical, social interaction, finance, motivation and personal. The data analysis revealed two influencing factors,

number of family members and the relationship with the lecturer, using Decision Tree algorithm with time and accuracy better than SVM and KNN algorithms [12].

Four machine learning algorithms were compared with 10 variables randomly sampled. The results show that Random Forest did well comparing to Neural Networks, Support Vector Machines and Logistic Regression in terms of the correctly predicted dropouts and sensitivity. It has been proved that the prediction is better when the students' current data are used to train the model than with previous data [13].

The authors used the classification trees techniques (CART and QUEST) to predict students at risk of dropout. The predictor variables were the students' socio-demographic and academic data. Three models have been developed, the induction-week model, the first 6–7 weeks' model and the end first semester model. These models are named according to the time to extract the data. The results revealed that the predictor factors are only academic performance while studying in the program and not previous performance before entering the university [14].

Gradient Boosted Trees and Deep Learning evaluated with (CV) Cross validation and (AUC) Area Under Curve metrics gave the best results of under-graduated students dropout prediction. Special attention was given to the previous high school performance in feature selection and feature extraction steps to implement the models [15].

The authors focused on early detection of students at risk of dropout in order to make an intervention. The class labels to predict by the machine learning algorithms are Pass, Fail, Conditional Fail, Repeat the Year and Repeat a Single Semester. The experiment is based on 32 features at the start point, but more features in other experiments were added progressively in the weeks after. The goal in this study was the detection of at risk students as soon as possible more than the search for the best accuracy; however, in less than 4 weeks it gave an accuracy of 97%. This model is ongoing and its data must be modified through time. It might not be applicable in other institutions or contexts as it is, but the idea of editable model that provide new results at any moment through the year still applicable [16].

The authors classified the student representations into three: Global Features-Based, Local Features-Based and Time Series with the appropriate learning algorithm for each of them. The experimental implementation showed that the Local Feature-Based was the best approach to predict the dropout, and the more the model is complex the more the computational costs increase [17].

The authors predict students' dropout in the case of programming courses using Online Judjes with Decision Tree algorithm. They used specific features related especially to the student's behavior and performance while trying to answer and submit programming codes such as number of student logins between the beginning and of a session [18].

Data of about 28000 students collected from 8 universities in Chile was subject of a study to identify high-risk students of leaving their studies. Two models are deployed through the use of Logistic Regression, one with data collected at the beginning of the semester; the second included the GPA of the first semester. The models were fit separately to each university. The low engagement and performance in the first weeks in online courses, attendance sheets and students' forum activity are signs that requires urgent intervention [19].

A case study of deep learning approach for predicting students' dropout at the University of Roma has been realized. The study concern 6000 students enrolled between 2009and 2014 and used Conventional Neural Network (CNN). The features were divided into three lists: the first and second one contains the administrative attributes of the students, and the third one includes the students' career attributes. Three models were implemented using CNN algorithm, and also Bayesian Networks just to compare and understand CNN. F1-measure was chosen as the adequate metric indicator into other evaluation metrics such as: Accuracy, Precision and Recall. The performance of the prediction has been improved from semester to another from about 67% for the first year's students to more than 94% for the third year students. That's why it is recommended to develop a real time and permanent predictive model [20].

The authors noticed that the dropout prediction in online higher education is basically a sequence labeling or time series prediction problem. They extracted seven behavioral features considered relevant the literature and added five others like Resource, Forum, and Subpage. They used the classifiers LR, SVM, RF, and DT and different evaluation metrics. The results classify LR as the best in prediction in terms of AUC and Accuracy [21].

A comparison between Logistic Regression and Decision Tree algorithms students' dropout has been done in order to predict university students' dropout. Both classifiers yield high accuracies with an advantage to decision tree. The data was collected from a higher education institute in Germany (KIT), and concentrate on different factors like performance in examinations, financial situation, motivation, health, and family issues. After the first semester the accuracy has reached 83%, after the third semester it has reached 95% [22].

In order be able to self-adjust results in an e-learning degree oriented program, the authors proposed a rule-based classification techniques JRip (JR), PART (PA), OneR (OR) and Ridor (RI) to stress the factors leading to the dropout of the students at different moments [23].

Three data mining classification techniques are used to provide university students dropout prediction. The algorithms are Machine Learning Based Decision Stump, NDTREE algorithms and Enhanced Machine Learning algorithms (EMLA). The data composed from 407 instances with 5 features and gave the following accuracies in order 78.3%, 70% and 30.7% [24].

The authors implemented an academic performance predictive model in private higher education institutions. They compared the results of the algorithms: Artificial Neural Networks (ANN), eXtreme GBoost, Linear Regression, Support Vector Machine, Naïve Bayes, and Random Forest. The models were fit with data composed of 10 features including marks, financial situation, study hours per week and English proficiency. The accuracy shows that ANN performed better than the five other algorithms in terms of classification of the students in two categories: those who will probably pass and those predicted to fail [25].

58% of the dropouts noticed have never failed a course, so there are other predictors. To identify them the authors employed three machine learning algorithms: Logistic Regression, Neural Networks and Decision Tree. They classify the dropouts without clear academic reasons as the existential dropouts, and analyzed the data of 45752 students between 2003 and 2015. The variables contain socio-economic status,

demographic, high school performance and other non-academic and non-academic predictors. The models have successfully predicted the graduates; however, the existential dropouts still a riddle that even non-academic pre-university performance predictors failed to foretell [26].

Data from 11 schools of a major university have been used to predict dropout of the first-year undergraduate students. The machine learning algorithms considered to choose the best one in terms of performance are Linear Discriminant Analysis (LDA), Support Vector Machines SVM and Random Forest (RF). The features vary between demographic previous academic performance and current performance. To evaluate the algorithms, the authors prioritized accuracy, sensibility and sensitivity than the metrics TP (true positive), TN (true negative), FP (false positive) and FN (false negative), because in this case it is about binary classification [27].

To predict wither a student will pass or fail and to evaluate the effectiveness of the predictive strategies in the literature, the authors used (ANN) Artificial Neural Networks and five other algorithms. ANN performed better and the results showed that the students' performance6 is correlated with features like group assignments and if the student is bursary or not [28].

2.2 Comment and syntax presentation

The different works of literature present the different ways of seeing the problem of students' dropout in the different institutions of higher education in the world. We can classify these works according to the algorithms, the type of variables used, or the time a student spent in the institution.

There are articles that have opted for a single Machine Learning algorithm to implement a model for predicting students' abandonment and performance. These kinds of studies sometimes focus on a single curriculum or different programs or even different universities to compare the effectiveness of the prediction among them as it is in [5], [6], [14], [19], [20], [27].

Other works have used several algorithms either to compare between them to choose the best [9], [12], [13], [26] or to combine them into one [4].

Online higher education or e-learning that is oriented to university degree is also concerned by several works as in [7], [18], [21], [23].

Some authors have considered that early detection plays an important role in ensuring timely action to avoid dropout as [9], [16].

There are dozens of variables used for training classification models, between those who have distinguished between local variables that are more specific to the program to which the student belongs, and global variables that do not take into account program data [17]. Others have used questions surveys to collect the data needed [11].

In general, we have noted in the literature the most widely used categories, and we list the variables belonging to each category in Table 1. They are divided into six classes: (A1) Academic features in the current study program, (A2) previous Academic features before joining the current program, (SD) Socio-Demographical features, (In) the features of the Institution and the program; (Bh) Behavioral characteristics and personality of the students in the program; (Fn) Financial features of the student.

Table 1. The features used in machine learning algorithms divided into categories

A1:	A2:	SD:	In:	Bh:	Fn:
– Access year – Academic performance rate – Success rate – Average mark – Mode round – Exam round – Educational level – English language literacy – Multiple choice test grade – Project grade – Lowest results – Highest results – Number of correct tests	– Access mark – Access form – Prior academic performance – Educational background before enrollment – Total failed courses – Total passed courses – Lowest previous assessment results – Highest previous assessment results – Average previous assessment results – Year of admission to the university – Semesters not enrolled in the past – Courses needed to graduate – Modules first time	– Marriage status – Sex – Birth – Family city – Country – Father's education level – Mother's education level – Residency – Working experience – Political status – Number of family members – Waiting time to study – Internet in the house – Health – Health Insurance – Home language – Family problems	– End degree – Mode round – Start year – End year – Degree – Certified program – Start course date – End course date – Number of unique days – Study program of enrollment – School shift – Inter-Itinerary – Pupil classroom ratio – Pupil teacher ratio – Years in system – Strict institution rules – Facility different for boys and girls	– Section activity – Project submission date – Number of click stream – Play video event – Interact with chapter – Study motivation – Suitability expectation to the majors – Relationship with students – Relationship with family – Relationship with lecturers – Order – Achievement – Autonomy – Change – Endurance – Aggression – Consistent – Heterosexuality – Number of chapter read – Forum Interaction – User Web access – Time used to submit answers – Addictions (Alcohol, Smoke, Pills, Solvents, Drugs),	– Student job – Father's job – Mother's job – Tuition fee source – Received a financial grant during the semester – Scholarship of the institution during the semester – Main source of household income

We identified the most commonly used machine learning algorithms in the literature, namely (LR) Logistic Regression, (DT) Decision Tree, (RF) Random Forest, (ANN) Artificial Neural Networks, (kNN) k-Nearest Neighbors, (SVM) Support Vector Machine, (NB) Naive Bayes, and (Oth1) the other less used algorithms like Linear regression, AdaBoost, Enhanced ML Algorithm, Linear Discriminant Analysis oreXtremeGBoost.

To evaluate the models, the authors used several evaluation metrics, which are the measures by which we can rate and understand the performance of a machine learning model. A quick way of visualizing the performance of a model is by a confusion matrix [29]. For our purposes, the "positive" class is considered the students who have graduated; and the "negative" class is the non-graduated students.

Table 2. General confusion matrix

Actual Positive	**True Positive (TP)**	**False Negative (FN) Type I error**
Actual Negative	**False Positive (FP) Type II error**	**True Negative (TN)**
	Predicted Positive	Predicted Negative

Accuracy (Acr): is the percentage of correctly classified samples. The accuracy of a machine learning classification algorithm is one way to measure how often the algorithm classifies a data point correctly. Accuracy is the number of correctly predicted data points out of all the data points. More formally, it is defined as the number of true positives and true negatives divided by the number of true positives, true negatives, false positives, and false negatives. A true positive or true negative is a data point that the algorithm correctly classified as true or false, respectively. A false positive or false negative, on the other hand, is a data point that the algorithm incorrectly classified. For example, if the algorithm classified a false data point as true, it would be a false positive. Often, accuracy is used along with precision and recall, which are other metrics that use various ratios of true / false positives / negatives. Together, these metrics provide a detailed look at how the algorithm is classifying data points.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1}$$

Precision and Recall (PR): are two numbers which together are used to evaluate the performance of classification or information retrieval systems. Precision is defined as the fraction of relevant instances among all retrieved instances. Recall, sometimes referred to as 'sensitivity', is the fraction of retrieved instances among all relevant instances. Perfect classifiers have precision and recall both equal to 1.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

It is often possible to calibrate the number of results returned by a model and improve precision at the expense of recall, or vice versa.

Specificity (Sp): describes the probability of the prediction being false when the actual class is false.

In simple terms, it describes how specific the model is when predicting negative instances. It is calculated as the ratio of true negatives to the actual negative cases. It can be calculated by using the values in the confusion matrix as the ratio of true negatives to the sum of true negatives and false positives.

$$\text{Sensitivity} = \frac{\text{TN}}{TN + FP} \tag{4}$$

F-Score (F-S): also called the F1-Score is a measure of a model's accuracy on a dataset. It is used to evaluate binary classification systems, which classify examples into "positive" or "negative".

The F-score is a way of combining the precision and recall of the model, and it is defined as the harmonic mean of the model's precision and recall.

The F-score is commonly used for evaluating information retrieval systems such as search engines, and also for many kinds of machine learning models, in particular in natural language processing.

It is possible to adjust the F-score to give more importance to precision over recall, or vice versa. Common adjusted F-scores are the F0.5-score and the F2-score, as well as the standard F1-score.

$$\text{F} - \text{Score} = \frac{2*\text{precision}*\text{recall}}{\text{precision} + \text{recall}} \tag{5}$$

Area Under Curve (AUC): In binary classification problems, the general rule of thumb is to use a probability threshold of 0.5 to make classification predictions. But for few scenarios, this threshold might not hold good and using a different threshold would be more appropriate. A Receiver Operating Characteristic (ROC) curve is the most commonly used method to visualize the performance of a binary classifier for different thresholds. It is obtained by plotting the True Positive Rate against the False Positive Rate. False positive rate is calculated as (1 - Specificity). From the ROC plot, we can calculate the Area Under the Curve (often referred to as simply the AUC) which is the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one.

Oth2: other less used evaluation metrics as Cross Validation or time of execution.

We present the essence of the literature review in Table 3 in two parts to facilitate the readability of the results. The columns indicate the articles in the references at the end of this article; the rows represent (F) Features, (MLA) Machine Learning Algorithms and (EM) Evaluation Metrics.

In the features columns we used the number of variables which belongs to each class category. When the details of the category are not indicated in the literature we make "X".

Table 3. Factors and tools involved in the prediction of dropout in the literature

Articles		[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]	[12]	[13]	[14]	[15]	[16]	[17]	[18]	[19]	[20]	[21]	[22]	[23]	[24]	[25]	[26]	[27]	[28]
F	A1	3	6	X	11	X	4	10	17	0	0	X	5	X	8	3	3	8	0	7	10	7	1	2	2	6
	A2	1	2	X	3	1	5	0	X	0	7	X	X	X	4	0	2	0	0	3	6	5	0	1	3	0
	SD	4	7	X	9	0	7	0	13	9	5	X	3	X	0	0	7	3	0	2	0	5	1	2	5	0
	In	0	5	0	1	0	2	2	0	0	7	X	3	X	0	0	2	1	0	0	0	6	2	2	3	1
	Bh	2	0	0	1	0	0	0	19	23	0	0	0	0	0	2	1	0	12	1	0	0	1	0	1	1
	Fn	0	3	0	1	0	0	0	X	5	0	0	0	0	0	0	1	0	0	2	0	3	0	0	1	2
MLA	LR		X			X	X				X							X	X						X	X
	DT			X	X		X	X		X		X	X		X		X	X	X	X			X		X	
	RF					X		X			X				X	X			X					X	X	X
	ANN	X			X		X				X		X				X					X				X
	kNN									X			X	X												X	
	SVM					X				X	X				X				X					X	X	
	NB				X	X			X				X		X										X	X
	Oth1	X				X	X						X						X		X	X	X	X		X
EM	Acr	X		X		X	X			X		X			X	X		X	X		X	X	X	X	X	X
	PR	X	X	X	X	X	X	X	X		X			X	X	X	X	X	X		X	X		X	X	X
	Sp	X									X			X				X			X			X		X
	F-S			X	X	X		X						X	X				X			X				X
	AUC			X									X	X					X	X						X
	Oth2							X		X			X								X					

3 Discussion

In the solutions proposed in the literature most of the studies are based on academic, socio-demographic, personal, behavioral financial and institutional data. Some of them consider that data about higher education degree oriented e-learning must be treated differently from face to face education. Given a program or an institution, the studies do not decide for the dropout and performance prediction the choice of the best features, the best algorithm and the best evaluation metrics that can be generalized for other similar programs or institutions. Indeed, in each case study, we find a different algorithm with different variables and different evaluation metrics.

The problem in the solutions proposed in the literature lies in the fact that there is no good model for all kind of higher education institutions, even at the same institution the results differ from one program to another. In each dropout prediction study the authors start from the beginning and manually test different machine learning algorithms with different training features. That costs time especially in the collection of data that we don't even know if it may have an impact on the dropout. The same university or the same country does not benefit from previous studies. One cannot generalize a model on all the other programs. When we talk about one algorithm is better than the others, this comparison is not absolute and it depends on the metrics we used to evaluate the algorithm.

To implement a machine learning algorithm with the goal: prediction students' dropout in higher educational context there are three challenges:

The first is forced by the availability of data, but using all the data we have is costly in terms of execution time and in terms of the difficulty of taking action after detecting the factors leading to dropout. In one side, the more variables are available in the outset, the more effective the study will be. In the other side, the more minimalist the model, the more generalized it can be on many institutions, even if it requires making some slight modifications.

The second challenge is the machine learning algorithm, it should be chosen in respect to the university recourses, data type, data size, time of execution authorized and more criteria.

The third level is the evaluation metrics that should judge the efficiency of the algorithm. There is no algorithm that is better than another always and in all metrics, so there is a decision to be done.

These challenges prove the multi-criteria aspect of the issue.

4 Recommendations

We raise the need for a global model (Figure 1) that takes into account the particularities of the study program and the data available in the concerned Higher Education Institution (HEI), in order to facilitate the choice of the best Machine Learning Algorithm (MLA) for the best available Variables (V) to consider, and the best Evaluation Metric (EM) to predict dropout as effectively as possible.

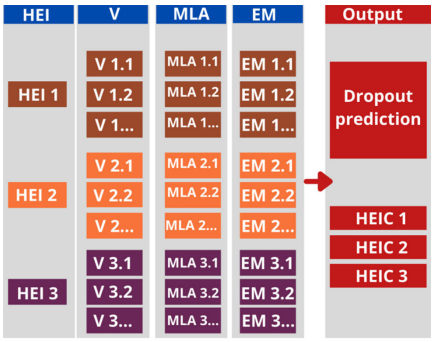

Fig. 1. Global model for higher education dropout prediction

To meet these needs, we put in recommendations the following actions:

a) Identification of Higher Education Institutions Classes (HEIC1, HEIC2, HEIC3…): Categories of institutions that have similarities in the prediction factors of the dropout. The objective is to be able to generalize the prediction models as much as possible in order to save time and effort, and to enable all institutions in the same category to benefit from a good model already successful in a similar institution.
b) Given the multi-criteria aspect of the problem identified in this work, a model of a (MCDM) Multi-Criteria Decision Making approach to choosing the three essential axes responsible for the prediction of the dropout must be the subject of next steps.
c) Development and testing of different Machine Learning Algorithms in a given higher education establishment in Morocco, then in different institutions and different programs and options, taking into account all the features of the literature to validate and judge the previous classification and model in a) and b).
d) Once all this work is done, the effort made can be exploited for predicting work integration. Because after successfully reducing the dropout rate, graduate students are faced with a new challenge, which is coming out of the unemployment phase.

5 Conclusion

Based on the literature review of the specific topic: dropout prediction in higher educational programs using Machine Learning techniques. We have classified the works according to the variables, algorithms and evaluation metrics used. The multi-criteria aspect of the problem has been proven. Given a higher educational institution, how to implement a machine learning algorithm to predict the dropout of student knowing that the larger the number of variables the more time we waste, but the performance of the algorithm would be better. In the other side if the number of variables is small, the algorithm can be generalized to other institutions in the same country, in the same city or at least in the same university.

The result of this work is a clear vision of research tracks that we have cited in the recommendations and that will serve all researchers who are interested in the quality of higher education, in the application of artificial intelligence or even to decision-makers

in the field of higher education, as a diagnostic and guide to getting to the point and solving one of the major problems of higher education worldwide which is students' dropout.

6 References

[1] Mukesh Kumar, A.J. Singh, & DishaHanda. (2017). Literature Survey on Educational Dropout Prediction. International Journal of Education and Management Engineering (IJEME), 7(2), pp. 8–19. https://doi.org/10.5815/ijeme.2017.02.02

[2] Mayra Alban & David Mauricio. (2019). Predicting University Dropout through Data Mining: A Systematic Literature. Indian Journal of Science and Technology, 12(4), pp. 1–12. https://doi.org/10.17485/ijst/2019/v12i4/139729

[3] Rahman Ali, Sungyoung Lee, & Tae Choong Chung. (2016). Accurate Multi-criteria Decision Making Methodology for Recommending Machine Learning Algorithm. Expert Systems with Applications. https://doi.org/10.1016/j.eswa.2016.11.034

[4] Ioanna Lykourentzou, Ioannis Giannoukos, Vassilis Nikolopoulos, George Mpardis, & Vassili Loumos. (2009). Dropout Prediction in E-learning Courses Through the Combination of Machine Learning Techniques. Computers & Education, 53(3), pp. 950–965. https://doi.org/10.1016/j.compedu.2009.05.010

[5] Francisco Araque, Concepción Roldán, & Alberto Salguero. (2009). Factors Influencing University Dropout Rates. Computers & Education, 53(3), pp. 563–574. https://doi.org/10.1016/j.compedu.2009.03.013

[6] Sulaiman Abdulsalam, Akinbowale Babatunde, Ronke Babatunde, & Moshood Hambali. (2015). Comparative Analysis of Decision Tree Algorithms for Predicting Undergraduate Students' Performance in Computer Programming. Journal of Advances in Scientific Research & Its Application (JASRA), 2, pp. 79–92.

[7] Mingjie Tan & Peiji Shao. (2015). Prediction of Student Dropout in E-Learning Program Through the Use of Machine Learning Method. International Journal of Emerging Technologies in Learning (iJET), 10(1), p. 11. https://doi.org/10.3991/ijet.v10i1.4189

[8] Sergi Rovira, Eloi Puertas, & Laura Igual. (2017). Data Driven System to Predict Academic Grades and Dropout. PLoS ONE, 12(2), p. e0171207. https://doi.org/10.1371/journal.pone.0171207

[9] Johannes Berens, Simon Oster, Kerstin Schneider, & Julian Burghoff. (2018). Early Detection of Students at Risk – Predicting Student Dropouts Using Administrative Student Data and Machine Learning Methods, Schumpeter Discussion Papers, No. 2018-006, University of Wuppertal, Schumpeter School of Business and Economics, Wuppertal, https://doi.org/10.2139/ssrn.3275433

[10] K. Limsathitwong, K. Tiwatthanont, & T. Yatsungnoen, "Dropout prediction system to reduce discontinue study rate of information technology students," 2018 5th International Conference on Business and Industrial Research (ICBIR), 2018, pp. 110–114, https://doi.org/10.1109/ICBIR.2018.8391176

[11] V. Hegde & P. P. Prageeth, "Higher education student dropout prediction and analysis through educational data mining," 2018 2nd International Conference on Inventive Systems and Control (ICISC), 2018, pp. 694–699, https://doi.org/10.1109/ICISC.2018.8398887

[12] T. Dharmawan, H. Ginardi, & A. Munif, "Dropout detection using non-academic data," 2018 4th International Conference on Science and Technology (ICST), 2018, https://doi.org/10.1109/ICSTC.2018.8528619

[13] M. Solis, T. Moreira, R. Gonzalez, T. Fernandez, & M. Hernandez, "Perspectives to predict dropout in university students with machine learning," 2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI), 2018, https://doi.org/10.1109/IWOBI.2018.8464191

[14] José María Ortiz-Lozano, Antonio Rua-Vieites, Paloma Bilbao-Calabuig, & MartíCasadesús-Fa. (2018). University Student Retention: Best Time and Data to Identify Undergraduate Students at Risk of Dropout. Innovations in Education and Teaching International, https://doi.org/10.1080/14703297.2018.1502090

[15] M. Nagy & R. Molontay, "Predicting dropout in higher education based on secondary school performance," 2018 IEEE 22nd International Conference on Intelligent Engineering Systems (INES), 2018, pp. 000389–000394, https://doi.org/10.1109/INES.2018.8523888

[16] Cameron C. Gray & Dave Perkins. (2019). Utilizing Early Engagement and Machine Learning to Predict Student Outcomes. Computers & Education, 131, pp. 22–32. https://doi.org/10.1016/j.compedu.2018.12.006

[17] Rubén Manrique, Bernardo Pereira Nunes, Olga Marino, Marco Antonio Casanova, & Terhi Nurmikko-Fuller. (2019). An Analysis of Student Representation, Representative Features and Classification Algorithms to Predict Degree Dropout. LAK19: Proceedings of the 9th International Conference on Learning Analytics & Knowledge, pp. 401–410. https://doi.org/10.1145/3303772.3303800

[18] Pereira F.D. et al. (2019). Early Dropout Prediction for Programming Courses Supported by Online Judges. In: Isotani S., Millán E., Ogan A., Hastings P., McLaren B., Luckin R. (eds) Artificial Intelligence in Education. AIED 2019. Lecture Notes in Computer Science, vol. 11626. Springer, Cham. https://doi.org/10.1007/978-3-030-23207-8_13

[19] Gad Yair, Nir Rotem, & Elad Shustak. (2020). The Riddle of the Existential Dropout: Lessons From an Institutional Study of Student Attrition. European Journal of Higher Education, 10(4), pp. 436–453. https://doi.org/10.1080/21568235.2020.1718518

[20] Lorenz Kemper, Gerrit Vorhoff, & Berthold U. Wigger. (2020). Predicting Student Dropout: A Machine Learning Approach. European Journal of Higher Education, 10(1), 28–47. https://doi.org/10.1080/21568235.2020.1718520

[21] Ahmed A. Mubarak, Han Cao & Weizhen Zhang. (2020). Prediction of Students' Early Dropout Based on Their Interaction Logs in Online Learning Environment. Interactive Learning Environments, https://doi.org/10.1080/10494820.2020.1727529

[22] Paul T. Von Hippela & Alvaro Hofflinger. (2020). The Data Revolution Comes to Higher Education: Identifying Students at Risk of Dropout in Chile. Journal of Higher Education Policy and Management, 43(1), pp. 2–23. https://doi.org/10.1080/1360080X.2020.1739800

[23] Clauirton Siebra, Ramon Santos, & Natasha Lino. (2020). A Self-Adjusting Approach for Temporal Dropout Prediction of E-Learning Students. International Journal of Distance Education Technologies, 18, pp. 19–33. https://doi.org/10.4018/IJDET.2020040102

[24] Francesco Agrusti, Mauro Mezzini, & Gianmarco Bonavolonta. (2020). Deep Learning Approach for Predicting University Dropout: A Case Study at Roma Tre University. Journal of E-Learning and Knowledge Society, 16, pp. 44–54.

[25] Sallan, G. and Behal, S. (2020). Prediction of Student Dropout Using Enhanced Machine Learning Algorithm. Advances in Mathematics: Scientific Journal, 9(6), pp. 3821–3826. https://doi.org/10.37418/amsj.9.6.61

[26] Francesca del Bonifro, Maurizio Gabbrielli, Giuseppe Lisanti, & Stefano Zingaro. (2020). Student Dropout Prediction. Artificial Intelligence in Education, pp. 129–140. https://doi.org/10.1007/978-3-030-52237-7_11

[27] Roderick Lottering, Robert Hans, & Manoj Lall. (2020). A Machine Learning Approach to Identifying Students at Risk of Dropout: A Case Study. International Journal of Advanced Computer Science and Applications, 11(10). https://doi.org/10.14569/IJACSA.2020.0111052

[28] Francis Makombe & Manoj Lall. (2020). A Predictive Model for the Determination of Academic Performance in Private Higher Education Institutions. International Journal of Advanced Computer Science and Applications, 11(9). https://doi.org/10.14569/IJACSA.2020.0110949

[29] Ting K.M. (2011). Confusion Matrix. In: Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-30164-8_157 [Accessed: December 10, 2021]

7 Authors

Khalid Oqaidi is a PhD student, IT engineer, and founder of a startup in the field of data science and market research and a company in computer science engineering (email: khalid.oqaidi@gmail.com).

Sarah Aouhassi is a professor engineer at the University Hassan II Casablanca, ENSAD Institute with a doctorate in Information Systems and a state engineering diploma from the national institute of statistics and applied economy (INSEA). Having experience of ten years within Hassan II University in Casablanca as a statistic chief officer. Her research focuses on the quality of Information System and its interaction with the quality of higher education. (email: haouhassi@gmail.com).

Khalifa Mansouri is a professor of computer science and researcher at the University Hassan II Casablanca, ENSET Institute. His research is focused on Real Time Systems, Information Systems, e-Learning Systems, Industrial Systems (Modeling, Optimization, and Numerical Computing). PhD (Calculation and optimization of structures) in 1994 to Mohammed V University in Rabat, HDR in 2010 and PhD (Computer Science) in 2016 to Hassan II University in Casablanca. (email: khmansouri@hotmail.com).

Article submitted 2021-09-28. Resubmitted 2022-06-22. Final acceptance 2022-07-22. Final version published as submitted by the authors.