GENERATING GLOBAL MODEL TO PREDICT STUDENTS' DROPOUT IN MOROCCAN HIGHER EDUCATIONAL INSTITUTIONS USING CLUSTERING

Khalid Oqaidi[1], Sarah Aouhassi[2] and 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*

ABSTRACT

The dropout of students is one of the major obstacles that ruin the improvement of higher education quality. To facilitate the study of students' dropout in Moroccan universities, this paper aims to establish a clustering approach model based on machine learning algorithms to determine Moroccan universities categories. Our objective in this article is to present a theoretical model capable of identifying higher education institutions that are similar in the dropout phenomenon. To avoid making Educational Data Mining Analysis on each higher educational programs predict students' performance, with such a classification we can reduce the number of studies to be done on one institution in each category of universities.

KEYWORDS

Students' Dropout, Higher Education, Machine Learning Prediction, Clustering.

1. INTRODUCTION

Students' dropout of higher educational institutions is an issue that gets more attention from the decision makers in the last years. In Morocco this problem is blatant especially in open access universities. The most common way to try to solve this problem is to predict students' dropout using machine learning. The prediction of "at risk students" allows the decision makers to take early actions before the abandonment. The problem that researchers face is that every dropout prediction study is different from other similar studies even in the same country or in the same university, which make it hard to an institution to benefit from other institutions actions. In this article we develop a global model that will automatically generate the best machine learning algorithm in dropout prediction our main contribution is providing a global model that will reduce time and give best results in higher educational dropout prediction in Moroccan institutions.

In section II we present recent related works that concentrate on higher educational dropout using machine learning algorithms and specifically clustering algorithms, in section III we comment these works and show that lacking points that lead to our contribution, in section IV we present our model, and in the section V we conclude with a summary and perspectives of our work.

2. RELATED WORKS

2.1 Moroccan Context

Students' dropout is one of the biggest problems that face the higher education in the world. In Morocco this issue is present by force, especially in open access institutions. In the report of (Higher Council of Education, Training and Scientific Research, 2018) it is stated that more than one quarter of bachelor students drop out only in their first year at the university.

We noticed the lack of a global model to predict students' dropout prediction using machine learning algorithms, and the need of grouping higher education institutions into groups depending on their dropout nature.

2.2 Surveys and Synthetic Studies

In (Kumar, 2017) the authors analyze different contributions of students' dropout prediction in India between 2009 and 2016, and present the results in a survey. They conclude that there are four Educational Data Mining categories: Classification, Clustering, Prediction and Association Rule mining. The most used machine learning classifiers are: Support Vector Machine, Decision Tree algorithms, Artificial Neural Networks, Logistic Regression, Naïve Bayes and Random Forest.

In a systematic literature done on 67 papers selected from 1681 ones the authors in (Alban, 2019) identified the techniques used in the literature to realize data pre-processing, the factors affecting the dropout, the techniques used to select these factors, the techniques used for prediction, their levels of reliability and the tools used.

A detailed review of 12 studies in educational data mining that use clustering algorithms is presented in (Dutt, 2015). The goal is to compare these studies according to their objectives, their algorithms, and their sources of data.

To classify students into natural groups depending on their characteristics, a review of clustering in (Iam-On, 2017) helped to develop a use case as practical guideline to facilitate detecting students at risk of dropping out.

Studying dropout in 160 Tunisian higher education institutions between 2013 and 2018 (Srairi, 2022) revealed the importance of contextual factors such as university accommodation in helping students to complete university education.

2.3 Use Cases

The authors in (Manrique, 2019) classified the representations of students into Global Features-Based, Local Features-Based and Time Series with the appropriate learning algorithm for each of them. The best approach to predict the dropout was the Local Feature.

To evaluate the efficiency of higher education institutions dropout in Brazilian universities, the authors in (Marilia, 2020) provide a comparative analysis between the three combined models in the first side: K-Means with Linear Regression, K-Means with Robust Regression, and K-Means with Support Vector Regression and the classic algorithms in the other side: Support Vector Regression, Bagging, Linear Regression, and Robust regression in the other side. The combined models gave satisfactory results in comparison with classic algorithms.

Another case study that uses clustering and decision tree in higher educational data mining is described in (Križanić, 2020). The clusters are selected here according to their similarities in learning behaviors.

The authors in (Kabók, 2017) study competiveness in higher education institution by clustering analysis to come up with groups of countries that share some similarities in this subject.

Making the Tinto's model (1973) operational in a way that it could be implemented by universities as an academic computational support system for predicting dropouts is the subject of (Nicoletti, 2019).

To detect students at risk of dropping out in German universities an Early Detection System (EDS) was developed in (Berens, 2019) by predicting their performance in an early stage using AdaBoost.

Our Model is divided in two parts: A clustering study of the higher education institutions based on institutional data "Figure 1", and a machine learning prediction study using students' data "Figure 2".

3. GLOBAL MODEL TO PREDICT STUDENTS' DROPOUT

Every higher education institution has its particular characteristics different from other institutions. There are differences relative to the programs, the students' background, the enrollment conditions, the administrative data, and the academic staff etc. That's why there are almost as many dropout prediction studies as the number of institutions. In our work we propose a model that brings together all dropout prediction use cases in one study.

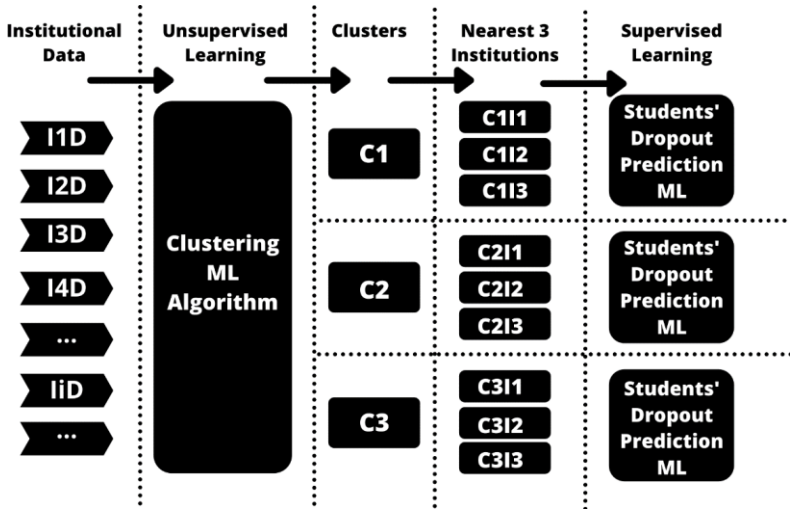

Figure 1. Identification of the nearest 3 institutions to the clusters' centers using the data of the institutions

3.1 Clustering the Higher Education Institutions

The goal of this step "Figure 1" is to come up with reduced number of clusters and figure out the best predicting algorithm of each cluster.

- First, we fit the machine learning algorithm with Data of the Institution j (IjD): information about the institution only. We gather all the available features such as: End degree, Start year, End year, Degree, Certified program, Start course date, End course date, Study program of enrollment, School shift, Pupil classroom ratio, Pupil teacher ratio etc. We can add all the details about the institutions that we can get access to.
- Then we use unsupervised machine learning algorithm to obtain institutions grouped by similarities.

We use the most frequent clustering machine learning algorithm in the literature: The K-Means Clustering (Mannor, 2011). We need to determine groups of institutions knowing only the dependent quantitative variables that concern the institutions. The qualitative variables must be converted to fit the model.

The first step is to choose a number of groups of institutions, and a number of arbitrary centers (institutions that represent the group) of the groups to initiate the algorithm. The algorithm decides for each center the closest institutions to build a cluster around. In each cluster the algorithm recalculates the distances to update the centers (it may be the Euclidian distance using the features' values). Now we repeat the process since we have new centers to redistribute the institutions on the clusters, so we update the institutions that belong each cluster. We obtain the same number of clusters, but this time they contain different institutions.

- Given one cluster i (Ci), we take three nearest institutions CiIj (CiI_1, CiI_2 and CiI_3) to the center of the cluster i. If the center itself is an institution we add two others, if it is fictive we add three.

3.2 The Best Machine Learning Algorithm per Cluster

- We use now supervised machine learning to predict the students' dropout in each CiIj "Figure 1".

To do so we gather all the students' data CiIjSD "Fig. 2" and we fit the most used machine learning algorithms in dropout prediction: Linear Regression, Decision Tree, Random Forest, Artificial Neural Networks, k Nearest Neighbors, Support Vector machine, and Naïve Bayes. Through the evaluation metrics in we sort the algorithms according to their performance; we retain the first one CiIjA.

- For each cluster we combine the three best algorithms CiIjA with AdaBoost as in (Berens, 2019) to come up with one machine learning prediction algorithm of the students' dropout per cluster CiA "Figure 2". Now we have for each cluster of higher education institutions in Morocco the most performing algorithm in term of students' dropout prediction.

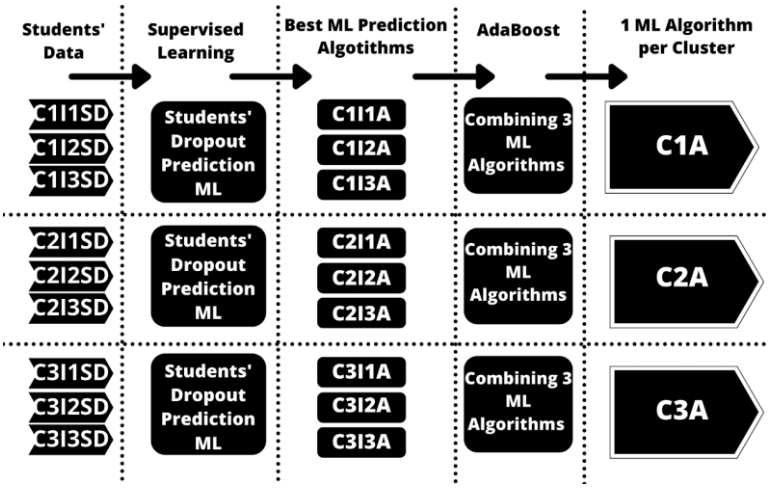

Figure 2. Generating the best dropout prediction algorithm per cluster using the students' data

3.3 Summary of the Model

Returning to "Figure 1" and "Figure 2" we started from the data of the institutions, which is by far smaller as students' data (it contains as many rows as the number of higher education institutions by a country). Then we apply clustering to reduce the number of institutions. We choose the three nearest institutions to the cluster's center by using Euclidian distance or equivalent. At this level we use machine learning to predict the dropout in each institution to come up with one algorithm per institution.

- To obtain the best algorithm per cluster we combine the three algorithms of the three institutions that represent the cluster (the nearest to the cluster's center). Now we have one algorithm per cluster.

3.4 How to Use the Resulting Model

When this model will be built in Morocco, given a new higher educational institution i (Ii) we collect just the data relative to the institution IiD "Figure 3", then we apply the first part of the model "Figure 1" to decide Cj. The model will generate automatically the best machine learning algorithm relative to this cluster CjA. We apply the second part of the model "Figure 2", we fit CjA with all the data available in this institution to predict students' dropout.

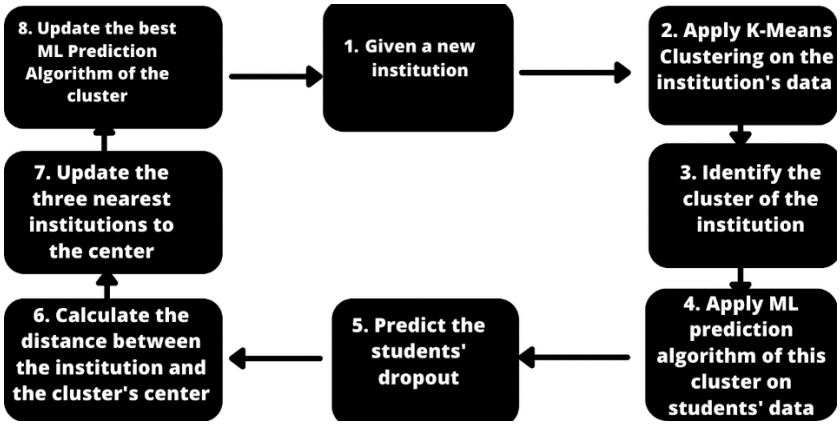

Figure 3. Use and auto-update of the model with new institutions

3.5 Model Application Scenario

According to (Ministry of Higher Education and Scientific Research of Morocco, 2020) there were 409 higher education institutions in Morocco in the season 2019-2020. We consider a scenario of 400 institutions.

- We consider 10 arbitrary institutions as initial centers of the clusters.
- We fit the K-Means Clustering algorithms with the institutional data of the 400 institutions.
- We have now 10 final centers, we take the three nearest institutions to each center including the centers if they are not fictive points. The result number is 30 institutions.
- We fit the supervised machine learning algorithms (Linear regression, Random Forest, Decision Tree…etc.) with all the available students' data to predict the dropout.
- After applying the evaluation metrics to these algorithms we obtain one best algorithm per institution, and in each cluster we have three best algorithms.
- To get the best algorithm per cluster we combine the three algorithms with AdaBoost to have just one performing better than each one CiA of the three. We have now 10 clusters with 10 machine learning dropout prediction algorithms: C1A, C2A… C10A.
- Given new institution in Morocco: number 401. We use just I401D, the model will determine the cluster of I401D. Let's suppose it is C7. Now we know that the best algorithm to predict the students' dropout in I401 is C7A. We fit C7A with the independent variables to predict the abandonment.

4. DISCUSSION

4.1 The Use Cases of the Model

Instead of trying to make a dropout prediction study in each higher education institution of the country, with such a model we make the machine learning prediction in less than 10%. The main contribution of this idea is to prepare a model for the researcher and the decision makers to not repeat the dropout prediction for any new institution even if it was not founded at the time of the development of the model.

This Model can be implemented in other countries to unify the dropout studies that use machine learning. The theoretical part still convenient even with a clustering algorithm different from K-Means and with different supervised machine learning algorithms.

4.2 The Model's Limitations

The main idea with this approach is that we do the effort once to make it easier for the studies that come after.

The first limitation is that institutional data must be available in all the institutions, and a convenient data must be available in all the institutions near to the clusters' centers.

The second limitation is the sustainability: once we decide the clusters for the country, new higher education institutions will be founded after. They may be with very different characteristics and far from any cluster. Considering the nearest cluster is not a perfect solution.

5. CONCLUSION AND FUTURE WORKS

Using machine learning to predict students' dropout in higher education institutions is largely used by researchers. Most of cases supervised machine learning algorithms are applied to tell in the output either a student will or not drop out. The problem of this approach is that every study stays local and the results concern only the institution in case.

To prepare a common ground on which all the dropout prediction studies in Morocco can start, we generated a global model based on clustering. The idea is to reduce the number of institutions to a smaller number of clusters, and then decide the best prediction supervised machine learning algorithm for each cluster.

Once we want to study the dropout in a given new institution we use the model to determine the cluster of this one, then we have already the best algorithm that can be used according to the model, so we fit it with the students' data to predict the abandonment.

Our next step is to implement this model in the Moroccan context to validate it. We will collect institutional data from all higher education institutions if it is available, otherwise we will collect the data of a representative sample of institutions. We have two objectives from that:

- Applying the model subject of this article to provide higher education decision makers in Morocco with groups of institutions based on some similarities.
- Deciding the best dropout prediction machine learning algorithm of each cluster. That can be obtained after collecting students' data from some institutions and applying our model.

Considering the second limitation of the model, we will work on the auto-correction of the model: once a new higher education institution is far from all the clusters we will affect it to the nearest cluster and update the cluster best algorithm. That will be subject of our future works.

REFERENCES

Alban, M. et al. (2019) "Predicting University Dropout Trough Data Mining: A systematic Literature," Indian journal of science and technology, 12(4), pp. 1–12. doi: 10.17485/ijst/2019/v12i4/139729.

Berens, J. et al. (2019) "Early detection of students at risk - predicting student dropouts using administrative student data from German universities and machine learning methods." doi: 10.5281/ZENODO.3594771.

Dutt, A. (2015) "Clustering algorithms applied in educational data mining," International journal of information and electronics engineering. doi: 10.7763/ijiee.2015.v5.513.

Higher Council of Education, Training and Scientific Research (2018) "L'enseignement supérieur au Maroc," Csefrs.ma. Available at: https://www.csefrs.ma/wp-content/uploads/2018/10/Rapport-Enseignement-sup--rieur-Fr-03-10.pdf, (Accessed: January 20, 2022).

Iam-On, N. and Boongoen, T. (2017) "Generating descriptive model for student dropout: a review of clustering approach," Human-centric computing and information sciences, 7(1). doi: 10.1186/s13673-016-0083-0.

Kabók, J., Radišić, S. and Kuzmanović, B. (2017) "Cluster analysis of higher-education competitiveness in selected European countries," Economic Research-Ekonomska Istraživanja, 30(1), pp. 845–857. doi: 10.1080/1331677x.2017.1305783.

Križanić, S. (2020) "Educational data mining using cluster analysis and decision tree technique: A case study," International journal of engineering business management, 12, p. 184797902090867. doi: 10.1177/1847979020908675.

Kumar, M. et al. (2017) "Literature survey on educational dropout prediction," International journal of education and management engineering, 7(2), pp. 8–19. doi: 10.5815/ijeme.2017.02.02.

Mannor, S. et al. (2011) "K-Means Clustering," in Encyclopedia of Machine Learning. Boston, MA: Springer US, pp. 563–564.

Manrique, R. et al. (2019) "An analysis of student representation, representative features and classification algorithms to predict degree dropout," in Proceedings of the 9th International Conference on Learning Analytics & Knowledge. New York, NY, USA: ACM.

Marilia et al. (2020) "A combined model based on clustering and regression to predicting school dropout in higher education institution," International journal of computer applications, 176(34), pp. 1–8. doi: 10.5120/ijca2020920396.

Ministry of Education, Vocational Training, Higher Education, and Scientific Research, (2020) " L'enseignement supérieur en chiffres," Gov.ma. Available at: https://www.enssup.gov.ma/storage/statistique/1.l'enseignement%20sup%C3%A9rieur%20en%20chiffre%20%202019-2020.pdf (Accessed: January 07, 2022).

Nicoletti, M. do C. (2019) "Revisiting the Tinto's theoretical dropout model," Higher education studies, 9(3), p. 52. doi: 10.5539/hes.v9n3p52.

Srairi, S. (2022) "An analysis of factors affecting student dropout: The case of Tunisian universities," International journal of educational reform, 31(2), pp. 168–186. doi: 10.1177/10567879211023123.