ISSN: 0067-2904

Predicting Student Dropout with Minimal Information

Nassim Mouchantaf , Maroun Chamoun
Doctoral School - Faculty of Engineering - Saint-Joseph university of Beirut, Beirut, Lebanon

Received: 8/4/2022 Accepted: 22/12/2022 Published: 30/10/2023

Abstract

Student dropout is a problem for both students and universities. However, in the crises that Lebanon is going through, it is becoming a serious financial problem for Lebanese private universities. To try to minimize it, it must be predicted in order to implement the appropriate actions. In this paper, a method to build the appropriate prediction system is presented. First, it generates a data source of predictor variables from student dataset collected from a faculty of economic sciences in Beirut between 2010 and 2020. Then, it will build a prediction model using data classification techniques based on identified predictor variables and validate it. Using open-source software and free cloud environments, a prediction program was developed. It consolidates, corrects, and normalizes the student's data. Then, it applies simple linear regression to show the correlation between the different variables and the student dropout, which allows us to select the factors that are highly correlated. From this point on, the program tries to predict the student dropout using different classification algorithms by machine learning on student dataset who left their courses either in success or in failure. Lastly, it measures the accuracy of the results and determines the best algorithm. In this study, the Artificial Neural Networks - Multilayer Perceptron showed an accuracy of 98.1% using only five variables. Finally, we evoke new avenues to further research and improve the model.

Keywords: deep learning algorithms, prediction system, student dropout.

1 Introduction

Normally, universities must continually adapt to the demands of globalization and the changing marketplace. This requires applying serious changes such as: permanent adaptation of training to the needs of the market, quality assurance of higher education and its accreditation, evaluation process at all levels, career management, professional insertion of graduates, research and its involvement with the market, application of information and communication technologies at all levels (teaching and administrative work), organization of seminars, colloquiums and congresses, international relations and conventions. All this requires educational institutions to have the financial and human resources necessary for their implementation, a situation that is not evident in some countries' situation.

Because the financial resources of universities are linked to the students' tuition, the universities must avoid losing students as much as possible. In Lebanon, there is no official figure on student dropouts, but according to a study with TechCARE (consortium of major Lebanese private universities, the Lebanese NREN), the dropout rate stood at 10.87% for the 2019-2020 academic year [1]. In addition to the financial loss, the universities lose reputation,

*Email: nassim.mouchantaf@gmail.com

notoriety and therefore their university ranking. The student also risks losing his future or at least his investment.

Traditionally, higher education institutions have not established evidence-based business practices. Increasingly, there has been a paradigm shift towards treating students as clients. This means recruiting the right students and maximizing student engagement by providing them with the best quality experience. This is a mutually beneficial goal, as the institution retains more tuition fees, and the students get the best quality experience during their studies. The challenge is to develop data processes that can identify students who are likely to drop out so that they can be allocated more resources to avoid this. The process, therefore, consists of identifying students at risk of dropping out as early as possible to take the appropriate measures to help them.

Current technologies based on deep learning algorithms [2] can implement simulation and prediction models of future situations by analyzing the data generated by the information system. In principle, after an analysis of existing data related to the specific topic, which is student's dropout, we can select useful information for a chosen deep learning algorithm to train the prediction system and thus help in decision making. Thus, in this study, the data generated by the grade management application of the faculty of economic sciences will be used to try to build the student's dropout system with the minimum amount of information.

2 Literature review

Machine learning is a method of data analysis where the systems can learn from data, identify patterns, and make decisions. Supervised learning is a set of machine learning tasks referring to algorithms that learn a function that maps inputs to outputs based on sampled input-output data pairs. They infer this function from labelled training data to make predictions that lead to decisions without being explicitly programmed to do so. Algorithms can discover patterns and establish relationships in the data that might be hidden from the average person. From this comes the need to build such models to apply to a sensitive topic such as student retention at university.

The interest of this project lies in classification, and more specifically binary classification, a task that requires the use of machine learning algorithms that learn how to assign a class label of two modalities to examples. The input data will consist of information describing each student, and the output label will represent their dropout status.

There is no good theory on how to map algorithms onto problem types; so it is generally recommended that a practitioner experiments and discovers which algorithm and algorithm configuration results in the best performance for a given classification task [3]. A large number of techniques have been developed for classification algorithms. Those applied in this study (with a preference for explainable algorithms) are the following:

Logistic Regression (LR) [4] is the go-to method for binary classification problems. It makes use of an equation as the representation of the output class. Input values are combined linearly using weights (estimated from the training data using maximum-likelihood estimation) to predict an output value being modeled as a binary value (0 or 1). The logistic function is an S-shaped curve that can take any real number and map it into a value between 0 and 1. This value would serve as a probability measure to classify data points, passing by a threshold value to separate the two classes.

Linear Discriminant Analysis (LDA) [5] was tested to address a few limitations that LR could have resulting from the nature of the data (being unstable with well separated classes and/or few examples for training). LDA does address these points and is the go-to method for multi-

class classification problems. Even with binary-classification problems, it is advisable to try both LR and LDA. LDA consists of statistical properties, namely the mean and variance of the data, calculated for each class. LDA also makes some simplifying assumptions about the data by assuming the variables follow a Gaussian distribution, and each attribute has the same variance.

Decision Trees (DTs) [6] are a powerful prediction method and are extremely popular because the final model is easy to understand by practitioners and domain experts alike. The final decision tree can explain why a prediction was made by tracking the "logical flow", making it attractive for operational use. A node in the graph represents an input variable, and the leaf nodes of the tree contain an output variable, which is used to make a prediction. Once created, a tree can be navigated following each branch decision until a final prediction is made.

The model representation for k-Nearest Neighbor (k-NN) [7] is the entire training dataset, making predictions using the training dataset directly. Because the entire dataset is stored, it is important to think carefully about the consistency of the training data. It might be prudent to update it as often as new data becomes available and remove outlier data. Predictions are made for a new instance by looking for the k most similar instances (or neighbors). In classification, the output label could be the mode (or most common) class value between those k neighbors. Determining the similarity of data points requires the implementation of a distance measure. As for breaking similarity ties, it is recommended to choose an odd k value for an even number of classes, and the inverse, to choose an even k value for an odd number of classes.

Support Vector Machines (SVMs) [8] is a high-performing algorithm with a little tuning. The Maximal-Margin Classifier is a hypothetical classifier that best explains how SVM works in practice. The input variables form an n-dimensional space. A hyperplane splitting this space is selected to best separate the points by their class (necessarily binary). In two-dimensions, this hyperplane can be seen as a line. The distance between the line and the closest data points (called support vectors) is referred to as the margin. The optimal line separating data points is the line that has the largest margin. The hyperplane is learned from the training data using an optimization procedure that maximizes the margin. This algorithm is thus time consuming with larger datasets.

Artificial Neural Networks (ANN) [9] are algorithms inspired by the nervous system intended to imitate how a human brain processes information. A neural network is an oriented graph, consisting of nodes, which in the biological analogy represent neurons, connected by weighted arcs. A neural network is essentially composed of the following three (3) layers: An Input Layer that takes in the raw information to be fed into the network, a Hidden Layer that applies transformations to the values fed to it, and finally an Output Layer that returns an output value corresponding to the prediction. A Multilayer Perceptron (MLP) is a class of feed-forward ANN that uses backward propagation to calculate through processing and correction arc weights.

In an article from Mckinsey [10], they discuss challenges and best practices deployed by higher-education institutions, such as involving good data hygiene, having standardized unified systems for processing all university data, and not letting "great be the enemy of good", as "it takes time to launch a successful analytics program".

Also, Jing Luan [11] examined the potential applications of data mining in higher education and explained how data mining saves resources while maximizing efficiency in academics.

3 Related works

The development of data mining methods and tools for analyzing data from educational institutions, defined as educational data mining (EDM), is relatively new in the field of data mining, as explained in studies conducted between 1995 and 2005 [12] and those beyond [13]. However, some claim the methodology is not yet transparent and it is unclear which data mining algorithms are preferable in this context [14]. The problems most often attracting the attention of researchers and pushing them to apply data mining at higher education institutions are focused mainly on the retention of students, improving institutional effectiveness, enrolment management, targeted marketing, and alumni management.

The implementation of predictive modelling for maximizing student recruitment and retention is discussed in many studies, however [14] focuses on predicting students' dropout.

Francesca Del Bonifro, Maurizio Gabbrielli, Giuseppe Lisanti & Stefano Pio Zingaro [15] have developed, using a few variables, a student dropout prediction model by trying several models Linear Discriminant Analysis (LDA), Support Vector Machine (SVM) and Random Forest (RF). With their basic functions, LDA and SVM gave good performance, while RF with the ALR function gave better performance. In short, all three gave satisfactory results with an accuracy of 85-90%.

J. Niyogisubizo, L. Liao, E. Nziyumva, E.Murwanashyaka and P-C. Nshimyumukiza [16] evaluated the effectiveness of the integrated model using Random Forest RF, Extreme Gradient Boosting XGBoost, Gradient Boosting GB, and Feedforward Neural Network FNN compared to the base models. After tuning the parameters, the results were compared based on training and testing results produced by FNN, RF, GB, XGBoost, and Stacking ensemble. It demonstrated that the integrated model achieved the best results in both training accuracy and testing accuracy,

In his article, Robert Reason [17] reviewed research conducted after 1990 related to the study of college student retention, examining research related to individual student demographic characteristics, highlighting the important features that take play in student retention.

P. Cortez and A. M. G. Silva [18] attempted to predict student failure by applying and comparing four data mining algorithms: decision trees, random forests, artificial neural networks and support vector machines. As for J-P Vandamme J-P, N Meskens and J-F Superby [19], they used decision trees, artificial neural networks and linear discriminant analysis for the early identification of three categories of students: low, medium and high-risk students.

D. Delen [20], in a comparative analysis of machine learning for student retention, applied four classification algorithms (artificial neural networks, decision trees, support vector machines and logistic regression) both individuals as well as ensembles. The results show the ensembles perform better than individual models. as a comparative analysis of machine learning techniques for student retention management.

In addition, at Nottingham Trent University (NTU), they conducted The HERE project [21] to improve student engagement and retention. They measured student engagement, which has proven to be an important indicator. They considered four factors:

- Campus attendance (tracked by building swipes)
- Library use
- Attendance and participation in classes
- Use of the university's online student portal.

Georgia State University has been a leader in leveraging data to provide individualized attention to students who need it most [22]. Over the past ten years, they have tracked over 140,000 student records and 2.5 million grades to identify 800 different factors that put students at risk of dropping out (i.e., Choosing the wrong course for their major and low grades in an introductory course required for the major, etc.). In doing so, they proved that SMS text messaging and Behavioral Intelligence are effective tools for driving student retention and persistence.

In their study titled "Data mining in higher education: university student dropout case study" [23] related to students who graduated from ALAQSA University between 2005 and 2011, Abu-Oda and El-Halees used a database that included all information about the students' academic history, and Decision Tree (DT), Naive Bayes (NB) as classifiers to predict the student dropout with an excellent accuracy.

In summary, all the methods and algorithms listed below have given reliable results, especially when they are well parameterized and well trained, and the best results were obtained when the solution integrates several algorithms. Since student practices and data differ from one university to another, this study attempted to simplify the solution while achieving excellent results, so that it can be adopted by most universities. This means minimizing the number of variables to be used and finding the best algorithm to go with it.

4 Review methodology

The goal lies in classification, specifically binary classification, a task that requires the use of machine learning algorithms that learn to assign a class label of two modalities to examples. The input data will consist of information describing each student, and the output label will represent their dropout status.

Several questions arise:

- What is the best machine learning classification model to rank students according to their performance, using a dataset, with a reasonable and significant rate of accuracy?
- What are the main key indicators that could help create the classification model to predict student dropouts?
- Could student dropouts be predicted with a reasonable and significant rate of accuracy using only the limited information available in the grade management application?

according to address those question, the study proceeded according to the following steps:

- Generation of a data source of predictor variables
- Identification of the varied factors that affect student performance
- Study of several machine learning algorithms
- Construction of a prediction model using the data classification techniques based on identified predictor variables.
- Measurement of results accuracy for each algorithm.
- Choice of the best one to be the student dropout prediction system.

The following diagram illustrates the steps to follow to develop the desired process.:

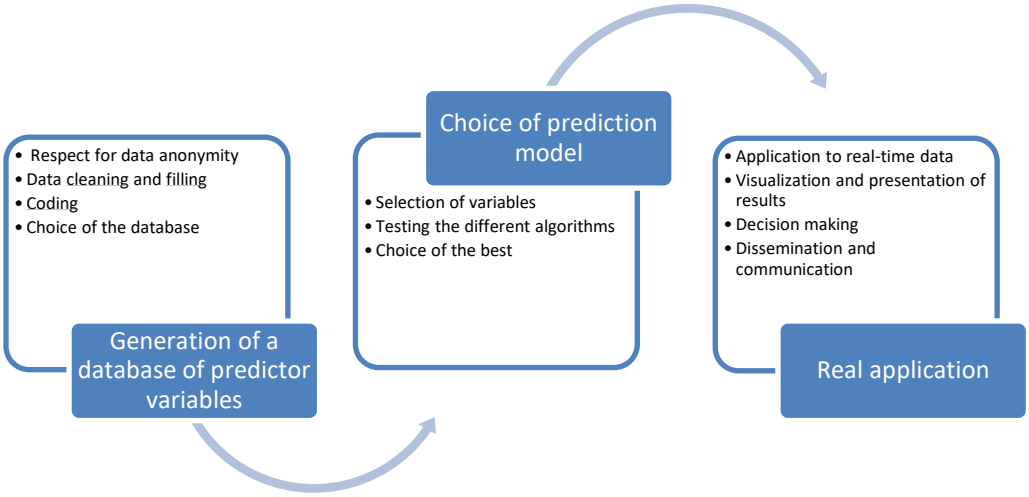


Figure 1: Process Architecture

4.1 *Generation of a database of predictor variables*

- Respect for data anonymity

To maintain the anonymity of the faculty and students, the data has therefore undergone a first treatment to meet the requirements of security and confidentiality of the faculty. The names and IDs of students and courses, the mail and other information identifying the student or the faculty were replaced by randomized characters.

- Data cleaning and filling

- The grade database is specific to a faculty, which enriches it according to its own specific provisions in relation to the general study regulations of the university (common rules), but also according to its own means and usage as well as its evolution in the time. This undoubtedly creates exceptions in the enrichment of the data thus generating data heterogeneity and therefore their purging is necessary.
- To minimize this heterogeneity of the data linked to the evolution over time of grades management practices, and to make the information consistent and correct, we cleaned up incorrect data and processed missing data:
- Choosing the most recent period, in our case the last ten (10) years.
- Delete all lines relating to teaching units that do not exist for various reasons.
- Calculating and filling the empty fields, such as age and Grade Point Average (GPA), where age = date of enrollment – date of birth, and the GPA value is obtained by a simple conversion of the decimal score according to a predefined table.

- Coding

Since the prediction models prefer to use numerical values, coding non numerical data should be done, such as for gender and place of birth. The value 0 of gender represents the male and 1 the female. As for place of birth, we created a table where each place is represented by a number.

- Choice of the database

The bibliographic research [24], [25] highlighted the role of some information in classifying struggling students. We therefore took what exists in the faculty database as well as those that can be added by calculation or deduction.

4.2 Choice of prediction model

- Selection of variables

Since the database contains a lot of information, it is therefore useful to select the minimum possible information that has the greatest influence on student dropout. By using simple linear regression analysis, the relationships between factors that are highly correlated with student dropout can be found.

- Testing the different algorithms

To find the models for creating the student's dropout prediction system, training and testing many classification models has to be done.

- Choice of the best

By calculating and comparing the accuracy of the obtained results, the best model can be chosen.

- Real application

After finding the most relevant variables and the most accurate algorithm for the student's dropout prediction, the university can apply this system on their real database and takes the corresponding decisions.

5 Finding

5.1 The database

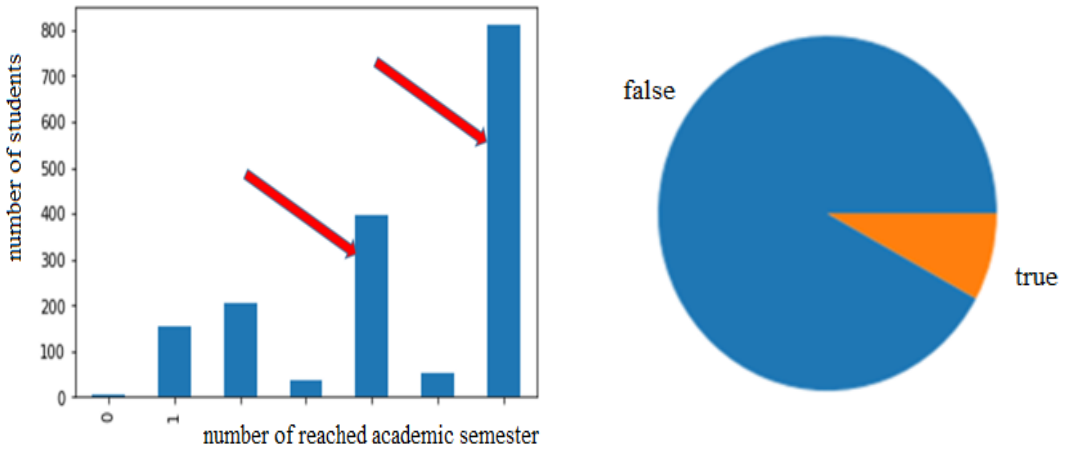
The new database becomes the following:

Table 1: number of items in the prediction system database

Number of years	total number of items	Number of items with the completed flag is yes
10	1273	728

Table 2: information in the prediction system database

Gender
Year of birth
Birthplace
Birth country
University program
Year of admission to university
Age of the student at the time of registration
GPA
Total university score: cumulative average
Dropout
Current study semester



(a) **(b)**

Figure 2: Distribution of students (a) by reached academic semester and (b) by dropout

The majority of students completed their academic courses, which is consistent with the pattern of dropouts. Bachelor's students have completed the required six semesters and Master's students have the required four semesters.

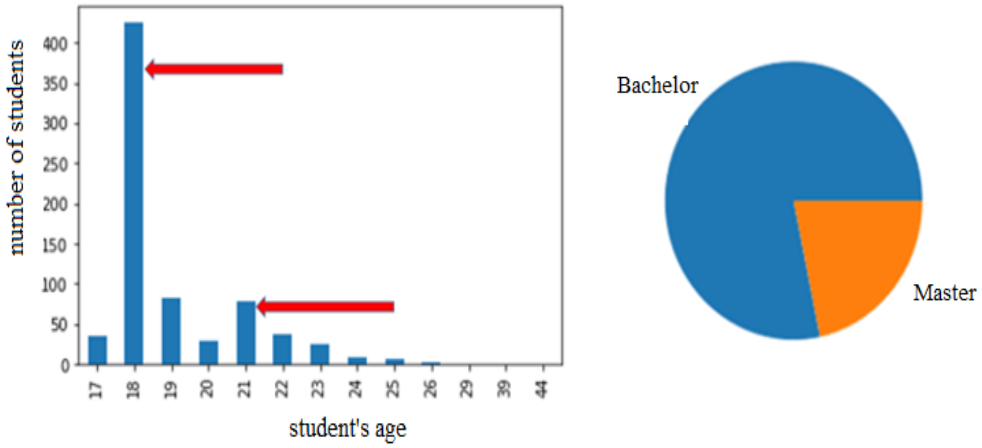
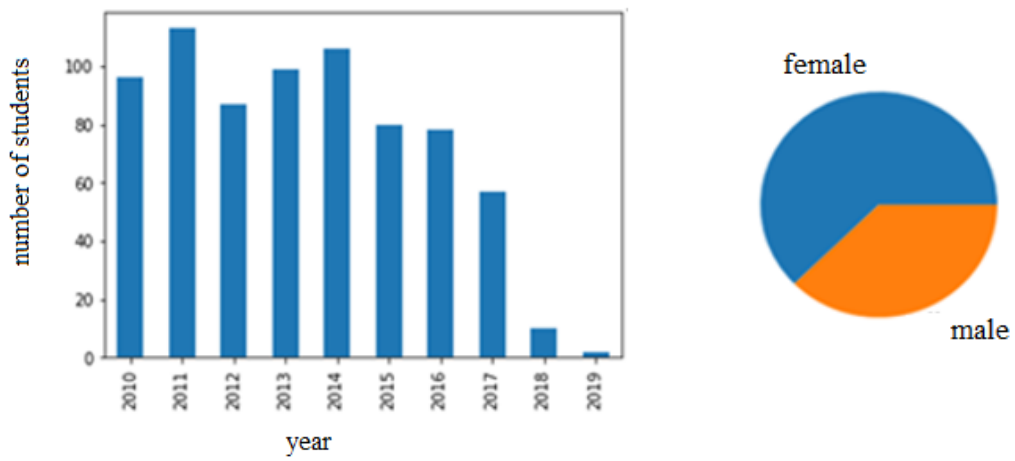


Figure 3: Distribution of students (a) per age and (b) per diploma

This shows that freshly graduated students (baccalaureate) of 18 years of age are enrolled in the first level of studies (bachelor's degree), while those of 21 years of age (baccalaureate + 3-year bachelor's degree) are enrolled in the master's degree.



(a) **(b)**

Figure 4: Distribution of students (a) by gender and (b) by number of students per year.

The distribution between females and males reflects the general characteristics of the country.

There has been a marked decline in the number of students enrolled in recent years, which may be due to the difficult times Lebanon is going through. The influx of Syrian refugees, the economic and financial crisis, the COVID-19 pandemic and the explosion of the port of Beirut have all put a strain on an already struggling educational system.

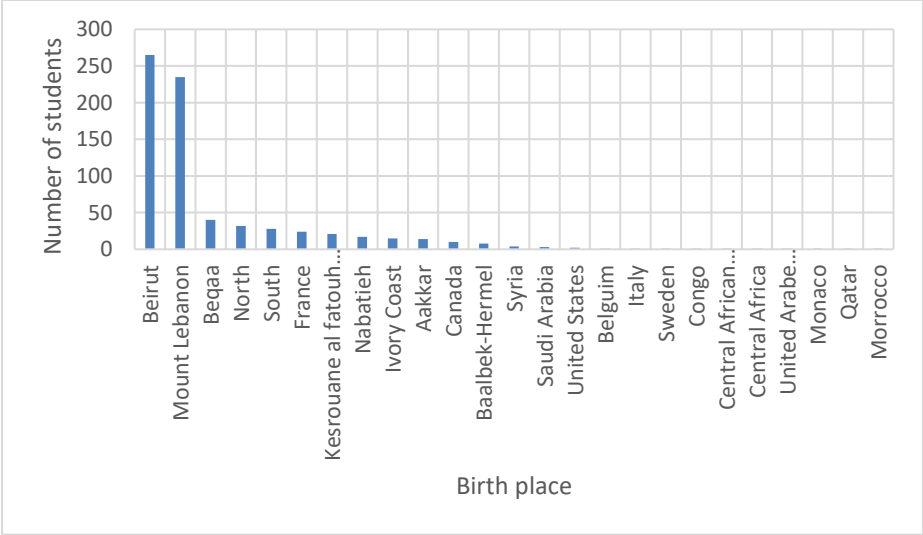


Figure 5: Distribution of students by birthplace Most students are of Lebanese origin and concentrated in Beirut and the nearby mountains.

5.2 Selected criteria

A correlation graph is useful to display the relationship between variables. Using simple linear regression analysis, the relationships between factors that are highly correlated with student dropout were found, as illustrated in the following graph:

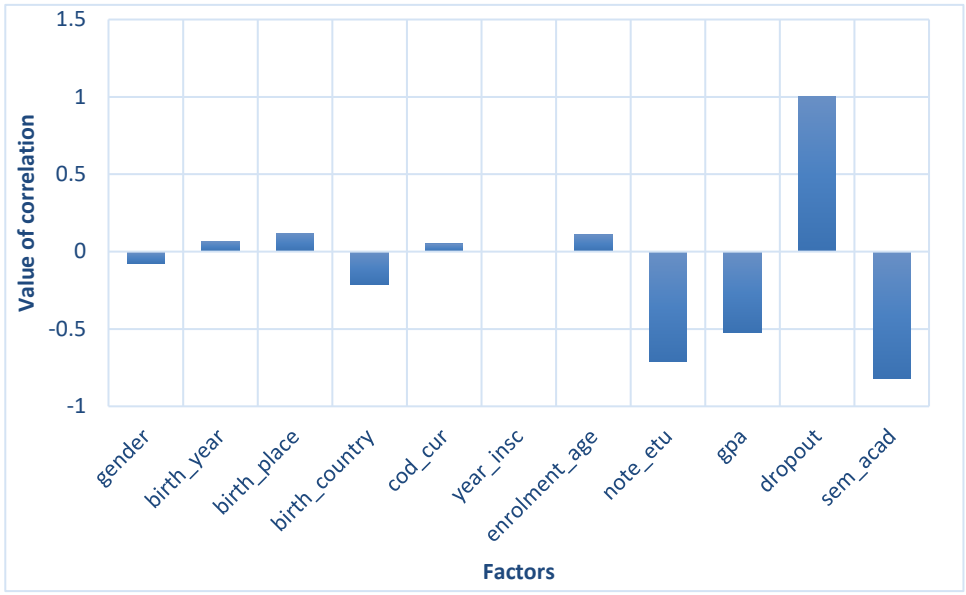


Figure 6: Shows the correlation for all variables with Dropout

As shown in the bar graph in Figure 5, the variables (student's cumulative average [note_etu], GPA, and current academic semester [sem_acad]) were negatively correlated with student dropout, coded as 1 for True and 0 for False. This means that the higher the student's grade or semester, the lower the value of dropout is likely to be, meaning the student will not drop out. It is not enough to study the correlation of variables with the output. It is necessary to check the correlation of all variables with each other; the independent variables could show potential dependencies with each other, which could lead to complications for the model. Below in Figure 6 is a triangular heat map showing the correlation of all variables:

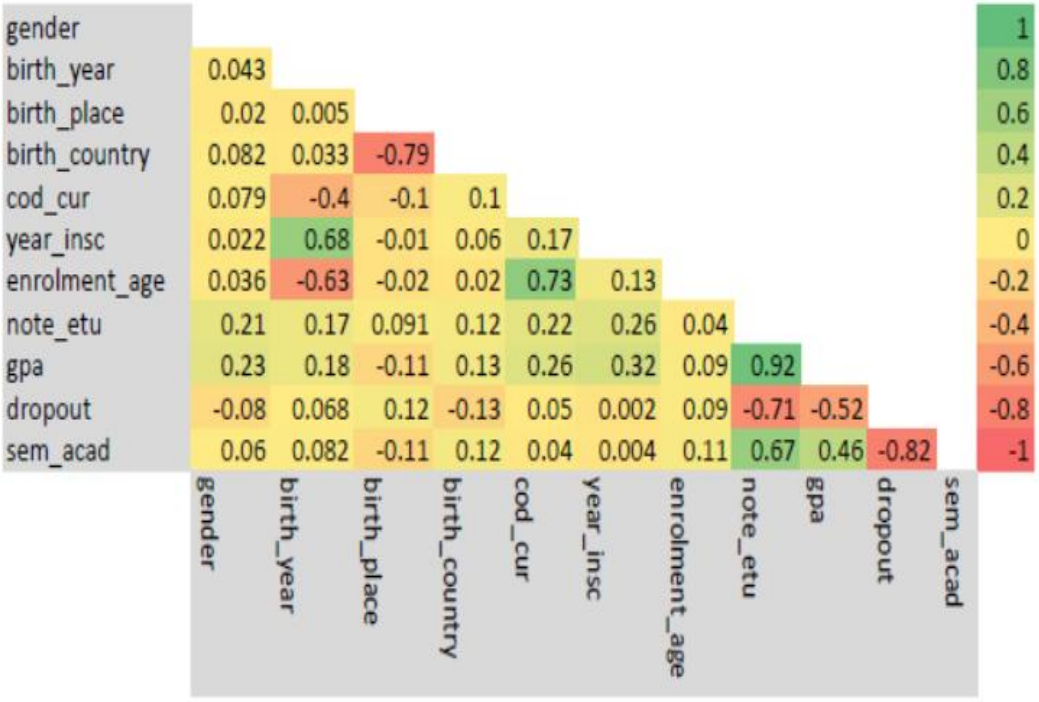


Figure 7: Triangular heat map shows the correlation between all variables

This not only shows the correlation found earlier, but also reveals that:
- Age of enrolment and year of birth are negatively correlated (-0.63): time passes and people's age as they pursue higher education; year of birth is constant over time.
- Year of enrolment and year of birth are positively correlated (0.68): students pursue higher education as they age.
- The age of enrolment and the program of study (curriculum) are strongly positively correlated (0.73): due to the encoding, the Master being encoded in 1 and the Bachelor in 0, and since one cannot pursue a Master without having completed one's Bachelor studies.
- The GPA and the student's cumulative average [note_etu] are strongly positively correlated (0.92): the higher the grade, the higher the GPA. The two variables express more or less the same idea since the cumulative average has more impact on the student's dropout (it has a higher correlation score), the GPA will be removed from the final set of variables.
- Current semester [sem_acad] and cumulative average [note_etu] are positively correlated (0.67): the further along in the semester, the more grades the students earn. This could reflect seriousness after pursuing a certain degree. It could also indicate that there are fewer dropouts as students' advance in their studies, or in other words, dropout would occur in the early years of study.
- The [gender] has a weak correlation with the dropout, and this is explained by the almost parity between boys and girls.
- The majority of the students being Lebanese, the [birth_country] variable will therefore not have a great influence on the dropout. On the other hand, [birth_place] reflects the students' level of education and therefore could have an influence on the dropout. The strong correlation between these two variables is related to the coding of the variable [birth_country] where 1 represents Lebanon and 0 represents abroad.

The final list of predictor variables can be limited to five elements:

Table 3: list of selected criteria

Birthplace
The university program
The current study semester
The total university score: cumulative average
The age of the student at the time of registration

5.3 Prediction model based on the identified predictor variables

Training multiple models and confidently choosing the best one is a guessing game. Cross-validation [26] comes to the rescue to estimate the competence of machine learning models. k-Fold cross-validation is a procedure with a single parameter - k - that refers to the number of groups into which a chosen sample of data should be divided (training and testing). The data are divided into k folds, and for each fold iteration, a different fold is chosen to test the model. The folds are performed keeping the percentage of samples for each class. The accuracy of the model is the average of the accuracy of the folds. A common practice in most predictive applications and a generally recommended number for k is 10. Cross-validation is useful for having many "shuffles" of the data, so the models will not be biased to a particular data distribution, which would lead to imperfect accuracy.

The correct evaluation metric also plays a role in evaluating the machine learning model. For a classification problem, the most popular evaluation measures are the classification accuracy and f1 score [27]. Precision is the ratio of the number of correct predictions to the total number

of predictions, while the f1 score is the harmonic mean between precision and recall, which are best explained through the confusion matrix displayed in the following table:

Table 4: Confusion matrix

	Predicted YES	**Predicted NO**
Actual YES	True Positive (TP)	False Negative (FN)
Actual NO	False Positive (FP)	True Negative (TN)

$$Precision = \frac{TP}{TP+FP} \quad (1)$$

the measure of the correctly identified positive cases from all the predicted positive cases

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

the measure of the correctly identified positive cases from all the actual positive cases
A way to find the "sweet spot" between the precision and recall is through the f1-score whose equation is:

$$2X\frac{Précision\ X\ Rappel}{Précision+Rappel} \quad (3)$$

Accuracy is used when the TPs and TNs are more important while the f1-score is used when the FNs and FPs are crucial. Failing to predict a dropout (FN) leads to failing to help a student in need. For that, the measure adopted is the f1-score.

The average accuracy reached to predict dropout for the models over a 10-fold cross validation is described in the following table (5). The best accuracy model in prediction is ANN with the five features retained by the RFE function.

Table 5: 10-fold cross validation results for all trained models

		Model					
		LR	**LDA**	**DT**	**K-NN**	**SVM**	**ANN**
ALL features	F1-score	80.8%	84.2%	82.5%	88%	76.9%	97.2%
	St. Dev.	10.8%	10.5%	12.2%	8%	12.5%	2.1%
	Range	70-91.6%	73.7-94.7%	70.3-94.7%	80-96%	64.4-89.4%	95.1-99.3%
5 features	F1-score	80.9%	85.3%	82%	90.2%	71.6%	98.1%
	St. Dev.	12.8%	11.2%	14.1%	5.1%	13.9%	1.5%
	Range	68.1-93.7%	74.1-96.5%	67.9-96.1%	85.1-95.3%	57.7-85.5%	96.6-99.6%

The following diagram shows the evolution of the learning and testing accuracy for the ANN model. Having a small gap at the end of the two curves is an indicator that no overfitting occurred during the training of the data:

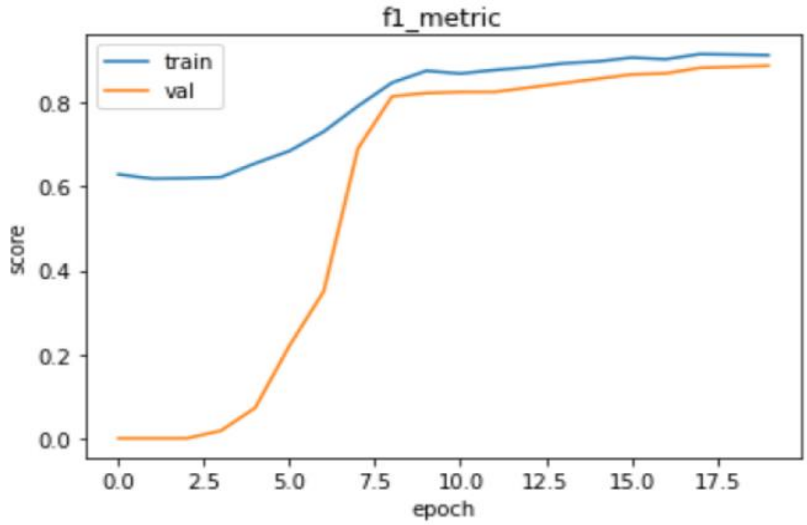


Figure 8: Graph showing the evolution of training and testing f1-score for ANN over epochs

6 Conclusion

A total of six machine learning algorithms were trained (LR, LDA, DT, k-NN, SVM, and ANN) and tested using k-fold cross-validation, and the accuracy of these models ranged from 70% to 99%. Ultimately, ANN was chosen to perform predictions on live data, as it far outperformed other machine learning algorithms trained to predict student dropout.

The ANN model may not be as good for other universities. It is, therefore, interesting to extend this exercise to all universities and find the best model, and then integrate the method into the university information system.

For universities that don't have a high percentage of student retention, the relationship between admissions and student retention needs to be studied. To do that, a database of admissions records large enough to be usable with the learning machines found is needed. The database may contain school of origin, grades and series of the applicant's baccalaureate. It is interesting to know the correlation between the school grades obtained and the dropout of students according to the specialty (by program). The statistics of the results obtained can be used to improve the orientation of the students.

The prediction model will also include characteristics related to the student's personal and financial situation, especially the value and nature of the financial aid offered by the university.

Finally, it is especially interesting to study not only the dropout of students but also their academic performance. The weakness of certain students can be predicted with respect to certain teaching units. The model will, therefore, have to take into account the following parameters: the teaching units, the first marks, the presence of the student in the course, his interactivity and the teacher. This could predict, from the beginning of the semester, the final result of the student in the concerned Course Unit UE. Therefore, actions can be taken to avoid the failure of students in some courses. And thus, avoid any discouragement of the student and his possible dropout.

7 Acknowledgements

I would like to express my gratitude to the dean of the faculty of economics for allowing me to use his faculty's data, to the IT department for providing me the corresponding database, and to the thesis director for his valuable support in bringing this project to fruition.

8 Disclosure and conflict of interest

The authors declare that they have no conflicts of interest.

9 References

[1] N. Mouchantaf, "The use of key performance indicators in the governance of Lebanese private universities.," Beirut, 2020.

[2] F. CHOLLET, Deep Learning with Python, Shelter Island: Manning Publications Co., 2018.

[3] J. Brownlee, "Types of Classification Tasks in Machine Learning," *Machine Learning Mastery*, 19 8 2020. [Online]. Available: https://machinelearningmastery.com/types-of-classification-in-machine-learning/.

[4] J. Brownlee, "Logistic Regression for Machine Learning," Machine Learning Mastery, 15 8 2020. [Online]. Available: https://machinelearningmastery.com/logistic-regression-for-machine-learning/.

[5] J. Brownlee, "Linear Discriminant Analysis for Machine Learning," Machine Learning Mastery, 15 8 2020. [Online]. Available: https://machinelearningmastery.com/linear-discriminant-analysis-for-machine-learning/..

[6] J. Brownlee, "How To Implement The Decision Tree Algorithm From Scratch In Python," *Machine Learning Mastery*, 11 12 2019. [Online]. Available: https://machinelearningmastery.com/implement-decision-tree-algorithm-scratch-python/..

[7] J. Brownlee, "K-Nearest Neighbors for Machine Learning," *Machine Learning Mastery*, 15 8 2020. [Online]. Available: https://machinelearningmastery.com/k-nearest-neighbors-for-machine-learning/..

[8] J. Brownlee, "Support Vector Machines for Machine Learning," Machine Learning Mastery, 15 8 2020. [Online]. Available: https://machinelearningmastery.com/support-vector-machines-for-machine-learning/..

[9] S. Sharma, "Artificial Neural Network (ANN) in Machine Learning," Data Science Central, 8 8 2017. [Online]. Available: https://www.datasciencecentral.com/artificial-neural-network-ann-in-machine-learning/.

[10] M. Krawitz, J. Law and S. Litman, "How higher-education institutions can transform themselves using advanced analytics," 8 8 2018. [Online]. Available: https://www.mckinsey.com/industries/education/our-insights/how-higher-education-institutions-can-transform-themselves-using-advanced-analytics.

[11] J. Luan, "*Data Mining and Its Applications in Higher Education," John Wiley & Sons, Inc.*, 16 4 2002. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/ir.35.

[12] Romero, C and Ventura, S, "Educational data mining: A survey from 1995 to 2005," 7 2007. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0957417406001266?via%3Dihub#!.

[13] R. Baker and K. Yacef, "The State of Educational Data Mining in 2009: A Review and Future Visions," *Journal of Educational Data Mining*, vol. 1, no. 1, pp. 3-16, 1 10 2009.

[14] G. Dekker, M. Pechenizkiy and J. Vleeshouwers, "Predicting Students Drop Out: A Case Study.," in *Journal of Educational Data Mining*, CORDOBA, SPAIN, 2009.

[15] F. Del Bonifro, M. Gabbrielli, G.Lisanti and S. Zingaro, "Student Dropout Prediction," in *AIED 2020: Artificial Intelligence in Education*, Ifrane, Morocco, 2020.

[16] J. Niyogisubizo, L. Liao, E. Nziyumva, E.Murwanashyaka and P-C. Nshimyumukiza , "Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization," *Computers and Education: Artificial Intelligence,* vol. 3, 2022.

[17] R. Reason, "Student Variables that Predict Retention: Recent Research and New Developments," *NASPA Journal,* vol. 46, no. 3, pp. 482-501, 20 10 2009.

[18] P. Cortez and A. Silva, "Using data mining to predict secondary school student performance," Semantic Scholar, 1 4 2008. [Online]. Available: https://www.semanticscholar.org/paper/Using-data-mining-to-predict-secondary-school-Cortez-Silva/61d468d5254730bbecf822c6b60d7d6595d9889c.

[19] J-P. Vandamme, N. Meskens and J-F. Superby, "Predicting Academic Performance by Data Mining Methods," *Education Economics,* vol. 15, no. 4, pp. 405-419, 26 10 2007.

[20] D. Delen, "A comparative analysis of machine learning techniques for student retention management," *Decision Support Systems,* vol. 49, no. 4, pp. 498-506, 11 2010.

[21] E. Foster, S. Lawther, C. Keenan, N. Bates, B. Colley and R. Lefever, "The HERE project toolkit: a resource for programme teams interested in improving student engagement and retention," Nottingham Trent University, Nottingham, 2012.

[22] T. Renick, L. Fifield, L. Page, H. Gehlbach and J. Lee, "How Georgia State University uses Behavioral Intelligence to improve student retention and persistence," Georgia State University, Atlanta, GA, 2019.

[23] G. Abu-Oda and A. EL-Halees, "Data mining in higher education: university student dropout case study," *International Journal of Data Mining & Knowledge Management Process,* vol. 5, no. 1, pp. 1-27, 2015.

[24] "leveraging big data predictive analytics for better student recruitment and retention," Higher Education Marketing, 15 2 2017. [Online]. Available: https://www.higher-education-marketing.com/blog/leveraging-big-data-predictive-analytics.

[25] A. Shahiri, W. Husain and N. Abdul Rashida, "ISICO 2015 : A Review on Predicting Student's Performance Using Data Mining Techniques," 2 11 2015. [Online]. Available: https://doi.org/10.1016/j.procs.2015.12.157.

[26] R. Ng, "Machine Learning - Cross Validation," 20 7 2021. [Online]. Available: https://www.ritchieng.com/machine-learning-cross-validation/.

[27] scikit-learn developers, "sklearn.metrics.f1_score," 2021. [Online].

[28] R. Wang, G. Harari, P. Hao, X. Zhou and A. Campbell, "SmartGPA: How Smartphones Can Assess and Predict Academic Performance of College Students," Ubicomp, Austin, 2015.