Student Desertion: What Is and How Can It Be Detected on Time?

Jonathan Vásquez and Jaime Miranda

1 Introduction

Student attrition—or student dropout—is understood as a student's failure in completing an educational program, which could be voluntary or involuntary. The first case means a student decides to stop to participate in the program through a formal quitting process according to the school's procedures, while the involuntary, the second case, refers to an institutional decision of finish the student's participation due to disciplinary or academic reasons [30]. The researchers from different disciplines such as Sociology, Psychology, Economy, History, Economy and recently, Data Mining have manifested interest in investigating student attrition phenomenon. Their motivations are related mainly to the desertion's costs, which can be identified from individual (frustration, financial debts, and future income reduction), institutional (funding reduction, opportunity costs, and low performance in indicators related to accreditation) and national perspectives (qualified worker reduction, benefit losses in the government investment and low performance in human development indicator) [29]. However, high complexity in the student attrition researches keep researchers and educational agents creating new methods and tools in order to reduce this global phenomenon [17], generating research opportunities for disciplines for contributing to the improvement of the management of the desertion.

A National Educational system might be considered as the engine that boosts economic and social growth due to its main objective of providing qualified workers, thus, an effective and efficient system on meeting this objective become important for any nation. Currently, Chile is working in a big change of its educational system.

J. Vásquez · J. Miranda (✉)
Department of Management Control and Information Systems, School of Economics and Business, University of Chile, Av. Diagonal Paraguay 257, 8330015 Santiago, Chile
e-mail: jmiranda@fen.uchile.cl

J. Vásquez
e-mail: jovasque@fen.uchile.cl

F. P. García Márquez and B. Lev (eds.), *Data Science and Digital Business*,
https://doi.org/10.1007/978-3-319-95651-0_13

Agents related to education like universities, professors, students, and organizations are participating in this process, due they are part of the system and any change will affect them. In this scene, implementing new mechanisms for reducing student attrition is coming considered as the important part of this reform. In fact, Education Minister started some investigation about desertion, and some of its results showed that student attrition is a non-minor problem for higher-education institutes. In fact, the Research Center of the minister, according to data analysis, estimates that 1 of 2 students leave the higher educational system once getting to enter. Apparently, the efficiency of the Chilean educational system creating qualified worker is similar to getting a face (or seal) when a coin is tossed [6].

In addition to the quantification of the problem of student desertion, it is important to understand why this phenomenon rises up, where, from the social sciences perspective, it means to identify the factors and predictors of desertion. Among the first ones, Pyke and Sheridan [23], from an econometric perspective, used logistical regressions obtaining as results that the financial and permanence time in the program are factors that positively influence in the retention. Ten years later, Sadler [26] applied econometric tools to explain desertion for freshman student of a nursing program. He identified that those students who showed a greater internal relationship with the nurse profession (being nurse) had a higher retention rate than those felt an external relation (doing nurse profession). Lately, thanks to the development of techniques and algorithms of Data Mining (DM) discipline, the application of these techniques in studies related to education started to grow, being the identification of predictors of desertion as one of the most important [22]. For example, Delen [8] identified, by the use of data mining techniques, predictors relate academic success as well as those that indicate if the student received funding such as scholarship or loan. In the Chilean case, research with an implementation of methods for student attrition reduction is focused to identify a different kind of student desertions and their costs [2, 9, 13]. For example, some universities have created academic support programs for the student previously they are admitted and once they are admitted to being part of the university. These methods approach on involuntary desertion—mainly academic reasons—, however, few methods, implemented by Chilean institutions, aim at voluntary dropouts. The challenge of conducting research that helps to manage student attrition becomes more important today, since Chilean higher education is in the process of structural changes, increasing free access to educational institutions, and therefore, more investment of the government funding by taxes of Chilean habitats. Therefore, the generation of new tools that improve the management of resources by reducing desertion will allow the investment to generate benefits for society, which would boost national development. In addition, educational institutions could improve their educational management, be fitting to high-quality educational standards reflected on national and international certifications.

2 Understanding Student Attrition

2.1 *Spady and Social Variables of the Individual Insertion*

Spady used the suicide principles of Durkheim [10]. These principles establish that the decision of suicide of any individual cannot be explained only by individual factors, but also this is a social phenomenon generated by the breakup of the person with his/her social system explained by the impossibility of integration to the society. Following this logic, Spady established that desertion would be a result of the no integration of the individual with his/her educational environment. In other words, he pointed the environment and family characteristics influence on student's expectations, and therefore, affect his or her social integration with the classmates and the final decision of leaving the program [27].

Figure 1 illustrates that family backgrounds impact directly on the academic potential and the attitude compatibility, interests, and the student's personal disposition to the characteristics of the environment. Both the academic potential and the normative conference affect the academic performance, the intellectual development and the integration of the student with the pairs. Each attribute that defines these factors impact directly to the social integration of the individual, which consequently will define the level of satisfaction hence the commitment with the educational institution. Therefore, every social factor related to family and their pairs will influence the final decision of the student's desertion.

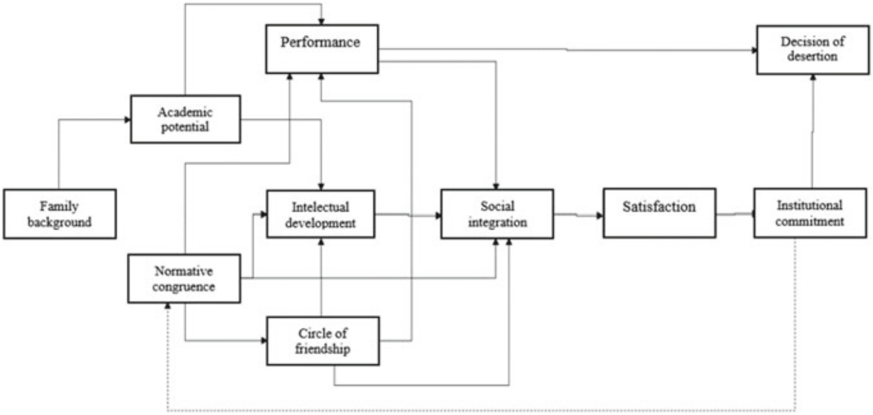


Fig. 1 Factors and their relation according to Spady

2.2 Tinto and the Socio-economic Factors

Tinto and Cullen [30, 31], added the exchange theory to the Spady's model. This theory postulates that human avoids any behavior that involves costs higher than benefits on the relationships, interactions, and emotional states generated by the interaction with their pairs and educational institution. These costs and benefits depend only on socio-economical characteristics of the individual. Under this theory, Tinto suggested that the students would stay in the program as long as the benefits received surpass the effort, dedication, and other personnel costs. Additionally, he established that the commitments of the student with the institution and their personal goals of professional formation are affected by their family background (e.g. socio-cultural level), their personal attributes (e.g. age and gender), and their pre-university academic experience (e.g. university selection exams performance). After a reasonable time enrolled in the program, the student will reevaluate his/her initial commitments according to their social integration and academic performance on the institution, which effects could trigger student desertion if the student notices that the costs are larger than the benefits. In short, Tinto proposed a rational behavior of the student in the continuous-evaluated decision of leaving or stay in the program.

2.3 Bean and the Organizational Factors

Bean [3] explained that the variables related to the student background, such as socio-economical, previous academic performance, and current residency, would influence the determinants of the student's relationship with the educational environment. In short, those students that have academic excellence records at the high school would get better performances at higher school, that it would increase the degree of satisfaction, institutional commitment, and, therefore, higher probabilities of no-desertion as the final decision.

Five years later, Bean joined Metzner [4] and extended the study to non-traditional students. In this new research, they postulated that the socio-demographic variables such as gender, age, and ethnicity are important to the heterogeneity of the body student, and this is important for any research about student desertion.

In summary, the variables identified as important for any study may vary according to institutional context, however, according to theory; we can establish that any research may consider at least three important groups: socio-economic, academic (both pre-university and university) and social (Figs. 2 and 3).

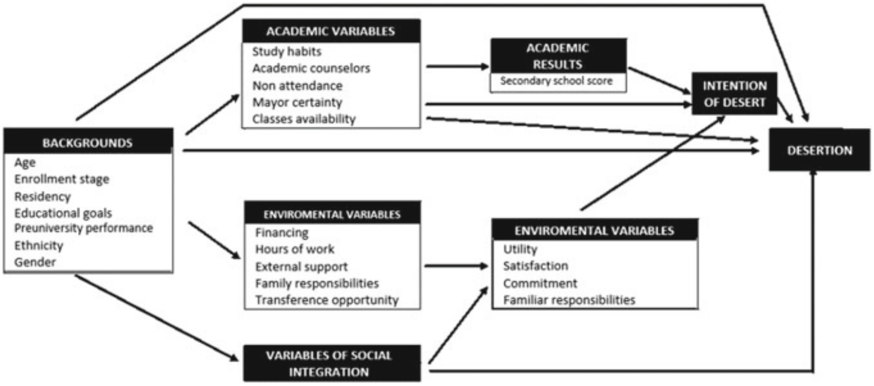


Fig. 2 The relation between the variables suggested by Bean

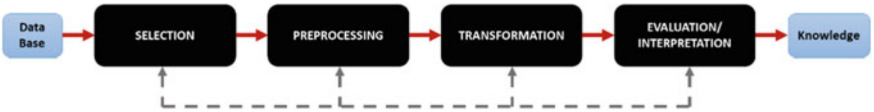


Fig. 3 Knowledge discovery in databases methodology suggested by Fayyad et al.

3 Literature Review

One of the objectives student dropout investigations is, amongst others, the identification of variables that would help to explain desertion, so they could be considered as predictors to detect early the student's decision of leaving the program. Econometric researchers have covered this objective by the identification of statistical significance of variables in the development of regression. For example, Pyke and Sheridan [23] used logistic regressions on a database of 601 postgraduate students, where academic, demographic, and funding were used as potential explanatory variables of desertion. The results showed that the greater the time the student is in the program and the funding awarded, the greater are the chances of graduating. In another research related to undergraduate nursing programs [26], the author compared two evaluation tools used for admission: Grade Point Averages (GPA) and a personal statement essay. According to the statistical analysis, the author found a difference statistically significant in the essay evaluations between deserting and non-deserting students, i.e., those students that completing the program tended to write in the essay their relationship with nursing in a more internal (being nurse) than external way (doing nurse profession). However, he did not find a statistically significant difference in the GPA, unlike other previously published studies did find it [5]. These investigations mostly have helped to identify explanatory variables of the desertion, opening a door for other disciplines, such as data mining, to contribute to the study the generation of predictive models.

The grow technologies' computation capability and reduction-storage costs have facilitated the implementation and development of new Data Mining techniques as solutions in any educational institutes. Indeed, researchers identified the challenge of using DM techniques tools in educational contexts, creating a new discipline named Educational Data Mining (EDM). This, considered as an emerging one, covers the development of techniques, methods, and models for the treatment of unique data for educational management. In the case of student attrition, researchers have faced three common problems [22]. One of them is the generalization of variables since the factors identified as most important by economists are not applicable in any educational context. The second problem is the definition of the temporality of desertion, because dropout reasons may vary in each semester. Finally, a third problem is the imbalance in databases.

3.1 Generalization of Variables

Yu et al. [33] applied three data mining techniques to identify desertion, suggesting that although many researchers had used parametric techniques, such as linear and logistic regression, this new perspective were new and allowed to detect non-linear and non-conventional relations between variables. In their research, features about hours transportation, residence (in or out campus), and ethnicity were considered the most crucial to predicting the early exit of a student. These variables differed from those found in econometric research, which indicated that academic performance at school was the most important predictor of attrition. Additionally, authors concluded variables identified by econometrical investigations were not generalizable for all educational context and called to the implementation of data mining techniques for future investigations. Following this tendency, Delen [8] identified that the variables considered as best predictors for the university context of its investigation were the academic success before and during university as well as the type of funding aid (scholarship or loan). In short, although the econometric studies allow initially to consider an initial set of variables, data mining techniques help to evaluate if these are applicable in different educational contexts, since as we saw previously.

3.2 Temporality and Unbalance

The concentration of desertion's occurrence is not always the same for all educational programs, due to the number of semesters and other specific features of the program (complexity, certification, and etcetera). In fact, it is possible to identify variations for the same kind of programs at amongst institutions from the same city, region, and country. This might imply that occurrence of dropout does not always focus on the same semester for all educational institutions, forcing researchers to identify when students drop out of the study program. For example, Alkhasawneh and Hargraves

[1] formulated a hybrid model with the aim of predicting retention for the first year; while Yu et al. [33] decided to do it for the second and third year. The effect of temporality also generates a difference among nations, since the programs' extension (in semesters) varies, even, amongst countries, being particularly longer in Latin America than in Europe or North America. In the other hand, in relation to the characteristics of the database, researchers have identified an imbalance in the data, being in general much less the number of students who deserted. This generates a problem in the data mining models, causing low precision for the desertion class and high for the retention. In short, it is important to evaluate and apply balancing techniques in order to reduce the problems associated with this kind of databases [8, 28].

3.3 Chilean Context

Díaz [9] did an extensive review of articles and publications related to student desertion. He concluded that there was a low volume of national investigations where data-mining techniques are applied in order to understand the desertion phenomena. In the same way, Himmel [17] claimed for a shortage of national investigations. Since then, study of student desertion has taken much more interest in educational institutions and researchers and they started to run a set investigations in order to understand student desertion and apply different techniques would support retention programs or actions to mitigate the costs of desertion [9, 20, 21].

4 Discovery of the Knowledge on the Database (KDD)

The authors Fayyad, Piatetsky-Shapiro, and Smith proposed a methodology named as Knowledge Discovery in Database (KDD) [11, 12]. They suggested a set of non-trivial activities with the objective of apply techniques correctly with analytic focus and then, increase the probability of projects success. Fayyad and his co-authors indicated that the basic issue approached by the KDD methodology is the data mapping of low level to identify knowledge and patterns; this would allow applications on different areas such as marketing, fraud detection, factoring, telecommunications, medicine, human resources, and education.

4.1 Selection

In this stage, the variables and observations are selected to be used in the data-mining project. The database is explored and analyzed with the goal of identifying variables that fulfill the following requirements: deterministic to the research; none or almost non-register error, be available in the future if the analysis is made again, and is measurable and collected on time according to the occurrence of the studied phenomena.

The researchers use statistical techniques and technological tools like automatic learning for support the selection of variables. Currently, there is a line of study that looks for generate technological methodologies for supporting variable selection. Reducing dimensions of the final data used in the data-mining project would help to solve the problem of capacity processing, improve the model's performances, and decrease the execution time of the algorithms.

4.2 Pre-processing

Once obtained the database with selected features, then all noise—missing values, outliers, etcetera, and bias must be eliminated. Noise problems can be solved through the replacement and reduction of the observations or variable. Depending on the type of data, the values can be replaced by mode (categorical variables), mean (numerical variables) or by the use of predictive models such as machine learnings. However, it is important to keep in mind that replacement could generate some bias, i.e., the analyst must preprocess carefully the variables.

4.3 Transformation

Some algorithms applied on the data mining stage require that the datasets meet some characteristics, such as only numerical variables. The transformation depends on the kind of variable selected and pre-processed on the previous stages. It is important to identify previously the existence of two kinds of variables (1) Numeric and (2) Categorical. The numeric variables imply that numbers, i.e., age and funding level, represents features and categorical variables capture any information by categories values, i.e., city, and gender. It is important to highlight that categorical variables are generally storage as text but can be eventually a number.

Some machine learning algorithms, as the case of Neural Net, require all variables be numeric and, for a better performance, normalized. The result of this stage generates a set of observations, with variables that are transformed into numbers and normalized according to data mining techniques' requirements.

4.4 Data Mining

The goal of this step is to extract patterns and knowledge previously unknown. The techniques applied can be of clustering, classification, or regression type. The last two are used mainly to build prediction models: in the case of classification, the tasks refer to the construction of models where observations are categorized on predefined labels; in the other hands, for regression case, activities refer to the use of models to predict variables related to a numerical indicator. The decision of implementing regression or categorization models depends crucially on the objective of data mining project.

4.4.1 Clustering

A clustering algorithm divides a set of observations X at n groups, and each of these groups is featured by a centroid. Currently, there are different clustering algorithms, being the most common k-means [15, 16]. This aims to divide M observations with N dimension into k clusters represented by a centroid. For each partition, the distance amongst observation to the centroid is minimal. Once the sets or partitions are identified, it is possible to create a model for each subset and so improve the predicting performance.

4.4.2 Imbalance Datasets

Imbalance database refers that proportions among different classes are not similar, which creates negative effects on the model's results and performance. The main problem is the model tends to learn more from the majority class than the minority one, or, in other words, the performance is much better in the majority than minority class. In order to solve this problem, it is necessary to evaluate the balancing technique application, i.e., Random Under-Sampling (selecting randomly observations of the majority class for removing from database until getting a balanced database amongst classes) and Random Over-Sampling (selecting randomly observations of minority class to adding the database until getting a balanced database amongst classes) [7].

4.4.3 Machines Learnings

Machines Learnings (ML) concerns to the study of programming computers (machines) in the order they can learn from datasets. ML is considered a branch of artificial intelligence and has been used in a variety of applications, such as text or document classifications, natural language processing, optical character recognition, and lately, educational context [19, 22]. Some examples of these machine learnings are Support Vector Machine, Decisions Tree, Neural Net and Logistic Regressions.

- Support Vector Machine (SVM): Support Vector Machine was introduced by Vapnik and Chervonenkis [32]. Nowadays, from machine learning discipline, SVM is a model of the supervised learning that bases on classification algorithms

and regression analysis. In other words, SVM classifies a set of points in the space by the hyperplane's partition and minimizes the cost of error of classification.
- Decision Trees (DT): Decision Trees is an algorithm considered suggested by Quinlan [24], based on the decision theories to make classification to the databases where algorithms are applied. Quinlan has made big contributions to the algorithms of decision trees, being the most known the C4 and ID3.
- Artificial Neural Net (ANN): ANN was initially introduced as a concept of a neural net by neurologists [18]. Fifteen years later, Rosenblatt [25] introduced the first simple perceptron based on biological neural net concepts, proposing fundaments of an artificial neural net (ANN). Nodes named as neurons compose an ANN, and each node receives a set of entries coming from other nodes and delivers outputs to others.
- Logistic Regression Logistic Regression is a special case of regressions used to predict the result of a categorical dependent variable. As an example, assume that the response variable y takes values 0 or 1, as in the case of desertion (dropout = 1 and not-dropout = 0). According to the postulated by logistic regression, the posterior probability of answer to the conditioned variable of the vector. After, the algorithm identifies the coefficients w of iterative form, usually through the method of maximum likelihood [14].

4.4.4 Classifiers

The learning machines deliver an indicator that shows the probability of a register belonging to a particular class. In some cases, depending on the algorithm, such indicator is reflected on the confidence, calculated by each observation, and can be used to determine classification thresholds, where to all confidences over the b threshold are classified to one class.

As an example, imagine that we have datasets of students and a machine learning is applied, so, after obtained prediction function, we obtain a confidence dropout class for each observation. On a predefined way, once the observations are ordered from the highest [1] to lowest (0) confidence, as a standard, observations with a confidence equal or bigger than 0.5 are classified as a dropout. This threshold could be more or less restrictive, increasing the classification threshold in the case that a bigger restriction is required to catalog a register as dropout and therefore less restriction to the other class no-dropout, or decrease the threshold if it is required to be less restrictive to the dropout class and more restrictive to the other one.

4.5 Interpretation and Evaluation

After Data Mining step, it is necessary to identify whether models are good or bad ones. In this step, evaluation and interpretation, the analyst must dominate the project context; due he/she should evaluate the results, and relate them to studied phenomena.

The performances of implemented models could vary for many factors, such as variables selected, the used algorithms, implementation of an appropriate optimization parameters process, among others. Therefore, it is imperative to use mechanisms to evaluate the performance of each technique in order to identify the best one. In this sense, literature has suggested different metrics to measure the predictive performance of the models. The most commons are the classification error and the accuracy. However, these metrics measure the general performance of models, assuming that classifying improperly any class imply the same errors, this means, have the same cost. Obviously, in some cases, like in the case of student attrition, to classify a student as dropout and finally stay, has not the same cost that classify the same student as not-dropout and finally leave. In short, the performance of the models can be evaluated based on its accuracy and classification cost. To measure both metrics it will be used the matrix of confusion, the tool usually used in classification applications, given the easiness of usage and quality of information delivered.

4.5.1 Confusion Matrix

The confusion matrix is a table that consists mainly of 2 rows and 2 columns—depending on classes number—, with information on the performance of the classification of a classification model. Usually, rows represent the instances predicted by the model, meanwhile, columns do the real instances observed. Additionally, the datasets are separated into two groups: train dataset, which is used for finding the function of the model, and test dataset, which is used for evaluation of the model's performance.

In case classification of two classes, these are named as positive and negative classes. A prediction of the class is obtained for each observation of test dataset by implementing the model generated through a learning process where train dataset was used. Each prediction is compared with the real class, and those observations predicted as positive and effectively were positive are denominated as True Positive (TP), but if they were not, they are evaluated as False Positive (FP). It is a similar case for the negative class, assigned as True Negative (TN) those observations predicted as negative and effectively were negatives, meanwhile, as False Negative (FN) the observations were positive but the model identified them as negative. The next table illustrates a typical confusion matrix (Table 1).

Table 1 Example of the confusion matrix

	True class		
Class prediction		+	−
	+	True positive	False positive
	−	False negative	True negative

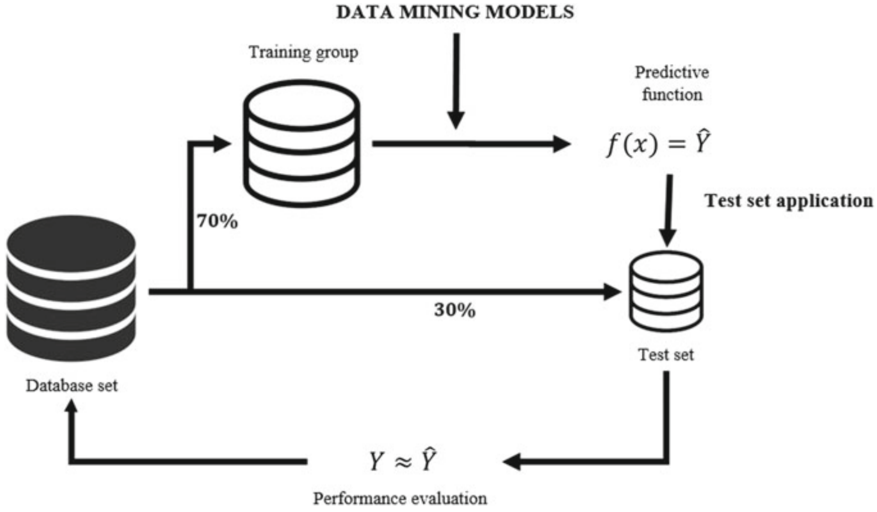


Fig. 4 Representation of cross-validation

4.5.2 Cross-Validation

Cross-Validation is a validation method mainly used to estimate how accurately a model will perform. As it was explained before, this method divided the datasets into two, train and test. This process is applied to k subsets of original datasets, everyone of equal size. One of these subsets is retained as the test data set. The remaining subsets are used as train data set, then the cross-validation method is repeated k times, with each of the k subsets used exactly once as the test data. The final performance evaluation is obtained from the k results of the k iterations. The performance is showed in averaged of these k results, so it is produced just one estimation of performance. The values most accepted for k are 5 or 10. Figure 4 illustrates a cross-validation method.

5 Experiment

5.1 FEN Case

The Facultad de Economía y Negocios (FEN, by its acronym in Spanish), it is the Business School of the University of Chile and it is located at the top ten on Latin-American rankings. The school manages four programs: (1) Commercial Engineering focused on Business (Business and Administration), (2) Commercial Engineering focused on Economics (Economics), (3) Audit, and Information Systems and Management Control Engineering. This study will focus on the last program.

Table 2 Distribution of desertion per entry groups

Year of entry	Non-desert (%)	Voluntary desert (%)
2007	66.0	34.0
2008	74.0	26.0
2009	74.0	26.0
2010	68.9	31.1
2011	59.6	40.4
2012	67.6	32.4
2013	79.8	20.2
2014	77.5	22.5
Total	70.9	29.1

In Chile, there is a national system for the application process to higher-education programs. This consists of a university selection test named University Selection Test (PSU, by its acronym in Spanish), that measures the knowledge of a student in four areas: Verbal, Mathematics, Sciences (Biology, Physics, and Chemistry) and History, Geography, and Social Sciences. Each test evaluates the knowledge of the students on a scale of 150–850 points. The attached universities to this process establish weights for each test, including the grades average obtained in high school, according to the program offered by each institution. The students, weighing the scores, apply by the national system to the programs, ordering these applications by preferences. All vacancies are completed by order of the score according to the weight established by the institutions.

It is important to highlight that each student can apply to up four programs and they are selected only in one. In the case of the Information Systems and Management Control Engineering program, the total of vacancies is at least 100 approx., varying each year according to evaluations and projections made by the school.

Approximately, in the case of Information Systems and Management Control Engineering program, 30% of the total of students that enter in a year, leaves voluntarily its studies, translating into an average of 31 spots lost every year. The variation per year of desertion it is showed in Table 2, where a 20 and 40% of a group of students with the same year of entry, quit the program voluntarily.

The six firsts semesters of the program are crucial to detecting voluntary dropout, because as shown in the Fig. 5, 97% of desertion occurs in the first three years, focusing on the third semester. In short, this study will focus on dropout that occurs during the first six semesters.

5.2 Databases

The dataset is collected from three databases: (1) Educational Administration System (SAD, by the acronym in Spanish), (2) Scholarships and Credits and (3) DEMRE

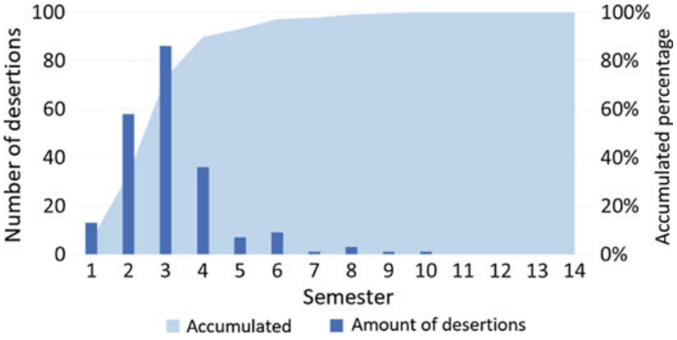


Fig. 5 Desertion behavior per semester. Own making

Table 3 Number of observations by semester

Semester	Desert	Non-desert	Total
Sem 1	13	595	608
Sem 2	58	591	649
Sem 3	86	591	677
Sem 4	36	578	614
Sem 5	7	486	493
Sem 6	9	577	586
General total	209	3,418	3,627

base. The Educational Administration System (SAD) stores and keeps students' information about enrollment, homologation of credits, academic performance, student's requests (such as temporarily stop studying and leave the program), and professor evaluation made by the students. In the other hand, information stored at the Scholarships and Credits is related to funding, as much in amount and funding type (scholarships and credits) for each student that receive each year funding such as Scholarships and Credits from the Ministry of Education and the internally from the University of Chile and FEN. Finally, the database sent annually by the Department of Evaluation, Measurement and Educational Register (DEMRE by the acronym in Spanish), managed by the University of Chile, has socio-demographic information pre-university academic performance, PSU scores and postulation of each student, given at the same moment of inscription for the PSU test.

The total number of observations was 3,627 with a distribution by semester showed in Table 3. The stored information in different databases allows obtaining 44 variables that cover socio-demographic, academic performances, and environmental and funding information of every student:

- Socio-demographic Variables: the number of family members, parent's educational level, number of parents alive, number of member working and studying at the different levels of the Chilean educational system, gender, and others related to work before entering into higher education.

- Performance and Academic Performance Variables: The variables related to academic performance of the student are stored in DEMRE and SAD systems. The data retrieved were academic performance in high school, score on each section of PSU, academic performance each semester (transcriptions and credits).
- Environmental Variables: Data from DEMRE system enable to obtain in-formation high school type (Technical or Scientific-Humanist), educational regime (Male, Female, or Co-educational high school), funding dependency type (Public Funding, the school is financed completely by the government and managed by municipalities; Private-Subsidized, the high school is co-financed by parents and government and managed by privates; Private, the institution does not receive funds from the government and is completely funded by students' parents and managed by privates), and application preferences. Additionally, from SAD systems, we enabled to obtain information related to postponements, voluntary participation on summer semester, and appreciation of the teacher's performance, which is considered as a proxy of satisfaction with the educational institution.
- Funding variables: related to family income, scholarships, and credits that each student receive every year.

5.3 Implementation

From the database, 6 datasets were obtained—each for semesters—, so we looked for six predictive model. We tested blending different techniques so we obtained 48 models for each semester, which generated 288 models in total. The following techniques were combined: (1) Clustering, (2) Unbalance Techniques (ROS, RUS and without unbalancing techniques), (3) Learning Machines (SVM, Neural Net, Decision Trees and Logistic Regression), and (4) Classification Threshold.

The software used was RapidMinner. In order to eliminate any bias and noise (such as missing values or outliers), only full observations were considered for the investigation, due to the low volume of data with these problems. Polynomial variables were transformed into binomial, generating n new columns where n was the number of different unique categories of the variable. From these n new columns, we selected $n - 1$ to avoid multicollinearity. Finally, the numeric attributes were normalized to a scale of 0–1 in order to match the range in every variable.

For clustering, we used the operator called X-means that balances the costs associated with precision and complexity of the model and delivers, as a result, the number of optimal centroids. Then, the operator assigns the observations to one of these centroids. This operator was tested in each semester dataset in order to obtain the best model.

Machine Learnings were applied to each cluster and to the complete dataset as well (without clusters), obtaining the performance of each one and identifying which of all was the best. Additionally, before applying Machine Learning algorithm, we tested balancing techniques (ROS and RUS) and evaluated how much was the improve compare to using imbalance datasets.

Finally, the classification threshold was applied in the testing process of the Machine Learning. Rapid Miner has operators that allow identifying the best threshold given the costs of classification errors to each class. Thus, after obtaining the fitted model in the test process, the confidence delivered by the algorithm is used as an input to the operator that identifies the best threshold, and then, the classifier uses this threshold and classify each observation in order to improve prediction performance.

6 Results

6.1 Best Models

In general, for every 6 semesters, the best models were composed of clustering, no-balancing techniques, SVM machine learning, and a classifier with an optimal threshold. Only in the case of semester 5 and 6, the machine learning in the best model was Logistic Regression.

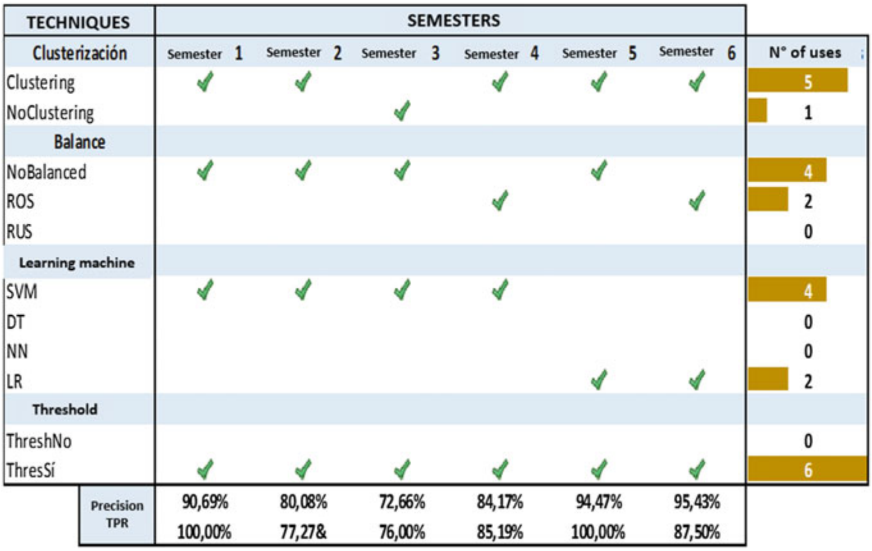
Comparing each technique, Logistic Regression and SVM machine learnings generated better performances to the models, as well as the classifier with an optimal threshold. In the case of the balance techniques, it was not clear its impact on the performance of the models, because is not always convenient its application.

6.2 Most Important Variables

Analyzing the best models of each semester according to the weights, the most important variables are those related to PSU, academic performance at higher school, professors rating, pre-university academic performance, parents' educational level, number of family members and how many are working, the application preference, participation on summer semesters and funding.

Considering those variables amongst the first quartile with the highest weight given by the models, the most important are PSU, specifically score on the verbal section, followed by the parents' educational level, and professor ratings. It seems that the student's background, mainly academic performance, must be strongly considered by the educational institution's manager of FEN, as well as the satisfaction of the student with the professors. In the other hand, extending the analysis to the second quartile, again, the most important variables were PSU performances, mainly on the tests of language and mathematics. It seems that the national university selection test is a good predictor of desertion (Fig. 6).

Analyzing Fig. 7, we identified that those variables related to family configuration and university performance are the most common amongst the 6 semesters. This is consistent with the models discussed by Spady [3], Tinto and Cullen [27], Bean [30],



TECHNIQUES	SEMESTERS						
Clusterización	Semester 1	Semester 2	Semester 3	Semester 4	Semester 5	Semester 6	N° of uses
Clustering	✓	✓		✓	✓	✓	5
NoClustering			✓				1
Balance							
NoBalanced	✓	✓	✓		✓		4
ROS				✓		✓	2
RUS							0
Learning machine							
SVM	✓	✓	✓	✓			4
DT							0
NN							0
LR					✓	✓	2
Threshold							
ThreshNo							0
ThresSí	✓	✓	✓	✓	✓	✓	6
Precision	90,69%	80,08%	72,66%	84,17%	94,47%	95,43%	
TPR	100,00%	77,27&	76,00%	85,19%	100,00%	87,50%	

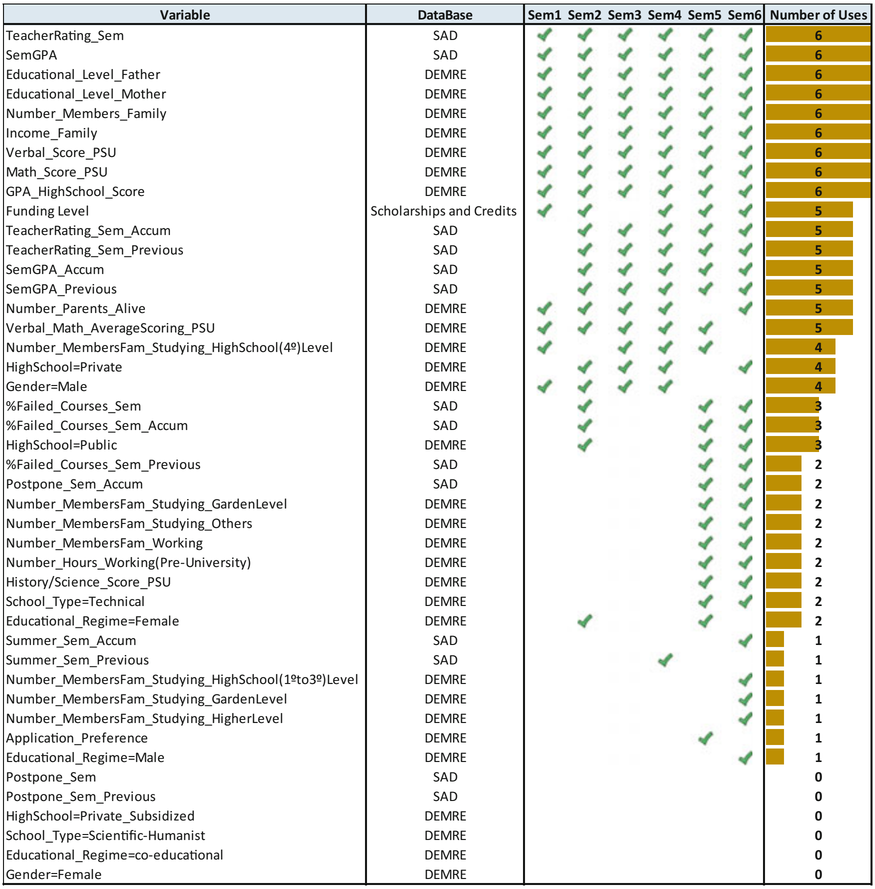
Fig. 6 Performance of the best models by semester

where they proposed that the family backgrounds and the academic performance are primary variables to explain the student desertion.

In summary, for each semester, it is important to evaluate PSU score of the student, followed by the educational level of the parents, university academic performance, funding, and family configuration in order to identify those students wanted to leave the program.

6.3 Deserters and Not-Deserters Profiles

According to the results, it was possible to identify the most important predictors for each semester. The most important for all semesters are the performance in PSU, the number parents alive, teachers rating, and university academic performance. However, the total set of predictors was different for all semesters, i.e., we identified a trend in the set of most important variables while advancing in the semesters. For the first four semesters, predictors related to the student's pre-university features were identified as the most important: specifically, PSU performance in all sections, number of parents alive, teacher rating, and the educational level of parents. In other words, as Spady [27], the variables related to academic potential (pre-university performance), family background, and educational context should impact the student's decision to remain in the program. Managers of educational programs must know these relationships amongst variables, and they could use it as support of selection process, as well as identify which students are potential deserters and implement



Variable	DataBase	Sem1	Sem2	Sem3	Sem4	Sem5	Sem6	Number of Uses
TeacherRating_Sem	SAD	✓	✓	✓	✓	✓	✓	6
SemGPA	SAD	✓	✓	✓	✓	✓	✓	6
Educational_Level_Father	DEMRE	✓	✓	✓	✓	✓	✓	6
Educational_Level_Mother	DEMRE	✓	✓	✓	✓	✓	✓	6
Number_Members_Family	DEMRE	✓	✓	✓	✓	✓	✓	6
Income_Family	DEMRE	✓	✓	✓	✓	✓	✓	6
Verbal_Score_PSU	DEMRE	✓	✓	✓	✓	✓	✓	6
Math_Score_PSU	DEMRE	✓	✓	✓	✓	✓	✓	6
GPA_HighSchool_Score	DEMRE	✓	✓	✓	✓	✓	✓	6
Funding Level	Scholarships and Credits	✓	✓		✓	✓	✓	5
TeacherRating_Sem_Accum	SAD		✓	✓	✓	✓	✓	5
TeacherRating_Sem_Previous	SAD		✓	✓	✓	✓	✓	5
SemGPA_Accum	SAD		✓	✓	✓	✓	✓	5
SemGPA_Previous	SAD		✓	✓	✓	✓	✓	5
Number_Parents_Alive	DEMRE	✓	✓	✓	✓		✓	5
Verbal_Math_AverageScoring_PSU	DEMRE	✓	✓	✓	✓	✓		5
Number_MembersFam_Studying_HighSchool(4º)Level	DEMRE	✓		✓	✓	✓		4
HighSchool=Private	DEMRE		✓	✓	✓		✓	4
Gender=Male	DEMRE	✓	✓	✓	✓			4
%Failed_Courses_Sem	SAD		✓			✓	✓	3
%Failed_Courses_Sem_Accum	SAD		✓			✓	✓	3
HighSchool=Public	DEMRE		✓			✓	✓	3
%Failed_Courses_Sem_Previous	SAD					✓	✓	2
Postpone_Sem_Accum	SAD					✓	✓	2
Number_MembersFam_Studying_GardenLevel	DEMRE					✓	✓	2
Number_MembersFam_Studying_Others	DEMRE					✓	✓	2
Number_MembersFam_Working	DEMRE					✓	✓	2
Number_Hours_Working(Pre-University)	DEMRE					✓	✓	2
History/Science_Score_PSU	DEMRE					✓	✓	2
School_Type=Technical	DEMRE					✓	✓	2
Educational_Regime=Female	DEMRE		✓			✓		2
Summer_Sem_Accum	SAD						✓	1
Summer_Sem_Previous	SAD				✓			1
Number_MembersFam_Studying_HighSchool(1ºto3º)Level	DEMRE						✓	1
Number_MembersFam_Studying_GardenLevel	DEMRE						✓	1
Number_MembersFam_Studying_HigherLevel	DEMRE						✓	1
Application_Preference	DEMRE					✓		1
Educational_Regime=Male	DEMRE						✓	1
Postpone_Sem	SAD							0
Postpone_Sem_Previous	SAD							0
HighSchool=Private_Subsidized	DEMRE							0
School_Type=Scientific-Humanist	DEMRE							0
Educational_Regime=co-educational	DEMRE							0
Gender=Female	DEMRE							0

Fig. 7 Variables used by best models for each semester

tools to prevent dropouts. Additionally, in the context of the researched program, we advise to extend the Academic Support Program (program that provides tutorials as a reinforcement) to cover more students for the first two years, due, according to results, academic performance of the first two years is considered as an important predictor for the semester five and six.

For the fifth semester, a strong change in the most important variable is identified, i.e., three of the most important variables were only features that are captured once the student entered the university, i.e., the same variables that are important for the first two semesters, are important for the fifth semester.

According to Spady [27], when the student enrolls in an educational program, according to his family and pre-university school background, he/she establishes his/her initial educational objectives and his/her institutional commitment. Later, after a sufficient time in which he/she interacts with its environment, these educational objectives and institutional commitment are adjusted, so they would trigger a desertion decision. This can be clearly seen in the variables most important by semester, where at the beginning of the program students with female gender, mostly from private-subsidized or municipal schools, with members of the family studying, and with high levels of funding, do not desert. Some of these characteristics, in general, are repeated in the three years, as it is the gender, members studying in another educational institution and the dependency group of the school.

The variables related to the participation of the student with their academic environment begin to take importance, mainly for the second year, where deserters are those students with low accumulated performance and provide a higher evaluation rate to their teachers.

The family configuration usually appears in all semesters, specifically, in desertions profiles. For example, as the higher educational level of at least one of the parents, the greater is the tendency to voluntarily desert by the student. The same happens when the number of members of the family studying in higher school increases. However, this is not the same when the members are studying at another educational level. Eventually, this could show the difficulty of the family in which two members are studying at the university.

7 Discussions and Conclusions

In general, a deserter profile can be described as those students who studied in private high schools, are male, their parents have higher educational levels, high family income, they have family members studying, they provide higher rating evaluation to their teachers until the third semester, and their performance at the PSU is slightly higher. In the case of no-deserters, their profile can be described as those who studied at private-subsidized high schools, are female, their parents have a lower educational level (less than a complete high school), the family incomes are relatively lower, receive higher funding, and PSU scores and high-school GPA are not very high.

The profiles previously mentioned would allow an early identification of students who would decide to stop studying in a specific semester. Additionally, given that academic performance variables are more important in all semesters, educational managers and advisor could pay attention to the academic performance of students, or, develop more supporting academic programs for the students.

As 5 of the 6 best models are composed of clustering techniques, it might be a sign that students tend to group and maybe, the characteristics of his/her group influence in the decision of desert. Indeed, the main variables that establish a difference amongst the groups are those related to family background and academic performance, variables directly related to deserter and non-deserter profiles.

The six models generated would allow the school to predict desertions in the first six semesters. These predictions can be used as an input in order to develop educational policies and reduce desertion rates, such as workshops of professional contextualization in order to improve institutional commitment and academic or psychological support programs in order to maintain personal objectives.

Acknowledgements Authors would like to appreciate the help provided in the data access for this investigation by Ariel La Paz, Director of Information Systems and Business School, Cesar Ortega, Chair of the Student Records Unit, and Marcia Oyarce, Social Assistant.

References

1. Alkhasawneh, R., & Hargraves, R. H. (2014). Developing a hybrid model to predict student first year retention in stem disciplines using machine learning techniques. *Journal of STEM Education: Innovations and Research, 15,* 35.
2. Barrios, A. (2013). Deserción universitaria en Chile: incidencia del financiamiento y otros factores asociados. Revistacis.
3. Bean, J. P. (1980). Dropouts and turnover: The synthesis and test of a causal model of student attrition. *Research in Higher Education, 12,* 155–187.
4. Bean, J. P., & Metzner, B. S. (1985). A conceptual model of nontraditional undergraduate student attrition. *Review of Educational Research, 55,* 485–540.
5. Byrd, G., Garza, C., & Nieswiadomy, R. (1999). Predictors of successful completion of a baccalaureate nursing program. *Nurse Education, 24,* 33–37.
6. Centros de Estudios MINEDUC. (2012). Serie Evidencias: Deserción en la educación superior en Chile.
7. Chawla, N. V., Japkowicz, N., & Kotcz, A. (2004). Special issue on learning from imbalanced data sets. *ACM SIGKDD Explorations Newsletter, 6,* 1–6.
8. Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems, 49,* 498–506. https://doi.org/10.1016/j.dss.2010.06.003.
9. Díaz, C. (2008). Modelo conceptual para la deserción estudiantil universitaria chilena. *Estud. Pedagógicos Valdivia, 34,* 65–86.
10. Durkheim, E. (1951). Suicide: A study in sociology (J.A. Spaulding & G. Simpson, Trans.). Glencoe IL Free Press. Work Publ. 1897.
11. Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine, 17,* 37.
12. Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM, 39,* 27–34.
13. González, L. E., & Uribe, D. (2002). Estimaciones sobre la "repitencia" y deserción en la educación superior chilena. Consideraciones sobre sus implicaciones. Rev. Calid. En Educ. Cons. Super. Educ. Diciembre Del 2002 (Vol. 77).
14. Han, J., Kamber, M., & Pei, J. (2011). Data mining: Concepts and techniques. Elsevier.
15. Hartigan, J. A., & Hartigan, J. A. (1975). *Clustering algorithms*. New York: Wiley.
16. Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 28,* 100–108.
17. Himmel, E. (2002). Modelos de análisis de la deserción estudiantil en la educación superior. *Calidad en la Educación, 17,* 91–107.
18. McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics, 5,* 115–133.
19. Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). Foundations of machine learning. MIT press.
20. Morales, F., Fuentes, R., Riquielme, S., & Kraemer, H. (2011). Impacto de la intervención del programa de inducción, adaptación y vinculacón a la vida universitaria en la facultad de ciencias empresariales de universidad del Bío Bío. Presented at the ENEFA (pp. 2730–2757).

21. Morales, F., Riquelme, S., Bascuñan, E., & Navarrete, M. (2014). Estudio sobre el éxito académico de estudiantes de ciencias empresariales de la Universidad del Bío-Bío. Presented at the ENEFA.
22. Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications, 41,* 1432–1462. https://doi.org/10.1016/j.eswa.2013.08.042.
23. Pyke, S. W., & Sheridan, P. M. (1993). Logistic regression analysis of graduate student retention. *Canadian Journal of Higher Education, 23,* 44–64.
24. Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning, 1,* 81–106.
25. Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review, 65,* 386.
26. Sadler, J. (2003). Effectiveness of student admission essays in identifying attrition. *Nurse Education Today, 23,* 620–627. https://doi.org/10.1016/S0260-6917(03)00112-6.
27. Spady, W. G. (1970). Dropouts from higher education: An interdisciplinary review and synthesis. *Interchange, 1,* 64–85.
28. Thammasiri, D., Delen, D., Meesad, P., & Kasap, N. (2014). A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. *Expert Systems with Applications, 41,* 321–330. https://doi.org/10.1016/j.eswa.2013.07.046.
29. Tinto, V. (2007). Taking student retention seriously. Syracuse University.
30. Tinto, V., & Cullen, J. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research, 45,* 89–125. https://doi.org/10.3102/00346543045001089.
31. Tinto, V., & Cullen, J. (1973). Dropout in higher education: A review and theoretical synthesis of recent research.
32. Vapnik, V., & Chervonenkis, A. (1964). *A note on one class of perceptrons* (p. 25). Remote Control: Autom.
33. Yu, C. H., DiGangi, S., Jannasch-Pennell, A., & Kaprolet, C. (2010). A data mining approach for identifying predictors of student retention from sophomore to junior year. *Journal of Data Science, 8,* 307–325.