Using machine learning methods to identify significant variables for the prediction of first-year Informatics Engineering students dropout

Felipe A. Bello*, Jacqueline Köhler†, Karen Hinrechsen‡, Víctor Araya†‡, Luciano Hidalgo† and José Luis Jara†
*Departamento de Ingenierías Multidisciplinares, Universidad de Santiago de Chile, Santiago, Chile
†Departamento de Ingeniería Informática, Universidad de Santiago de Chile, Santiago, Chile
‡Facultad de Ingeniería, Universidad de Santiago de Chile, Santiago, Chile
{felipe.bello, jacqueline.kohler, karen.hinrechsen, victor.arayas, luciano.hidalgo, joseluis.jara}@usach.cl

Abstract—**Student dropout is a phenomenon that affects all higher education institutions in Chile, with costs for people, institutions and the State. The reported retention rate of first year students for all Chilean universities was of 75%. Despite the extensive research and the implementation of various models to identify dropout causes and risk groups, few of them have been carried out in the Chilean higher education context.**

Our work attempts to identify, using machine learning methods, the variables with highest predictive value for student dropout by the end of the first year of study, within a 6-year Informatics Engineering programme with a rather high dropout rate of 21.9% reported on 2018. In that regard, we use the data of 4 cohorts of students (2012-2016) enrolled at the programme, to feed a random forest feature selection process. We later build a decision tree using the identified relevant features, which we later test using data of the 2017-2018 cohorts of students.

Despite the fact that the decision tree is over-fitted (97,21% training accuracy against 81.01% test accuracy), the process sheds light on the nature of the variables that determine whether or not a student remains at the end of their first year of study at the University. 6 of the identified factors are academic, and the remaining one is social-cultural.

Index Terms—**first-year student dropout, decision trees, random forest**

1. Introduction

Student dropout is a phenomenon where university students completely abandon their programme, school or department, faculty or university. In Chile, this problem affects all higher education institutions [21] with a wide variety of consequences, such as: (1) students who abandon their careers suffer psychological effects and generate a failure condition regarding their personal aspirations [17], which has an impact in their occupational path. (2) A social impact given by inequity and social differences. (3) Student dropout or career changes imply an additional cost to the institutions and to the State [19], since teaching efficiency indicators such as retention and graduation rates establish a credit structure to finance students [23] and student dropout has been included as an evaluation metric of higher education institutions [24]. Even though the Chilean state has broadened university coverage allowing a wide range of the population to study free-of-charge [20, 21], thus reducing the first year university students dropout rate from 29.5% on 2014 to 25.0% on 2018 [26]. However, the last reported rate is still considerable.

Within the Departamento de Ingeniería Informática (DIINF), Universidad de Santiago de Chile, the first year dropout rate for the 6-year engineering programme was of 28% on 2017 and of 21.9% on 2018, By the third year of study, the dropout rate increases almost to 50%, where most students abandon their studies during the third year, hence positioning as one of the departments with highest dropout rates within the Facultad de Ingeniería [30, 11].

Student dropout has been broadly studied during the last decades. Several factors have influence on the student dropout phenomenon: psychological, sociological, organisational, economic, interaction with the environment, perseverance, motivation, academic achievement and commitment, among others [1, 2, 4, 14]. Diverse models have been proposed to predict student dropout. Fishbein & Ajzen [18] consider personal intentions and beliefs; Bean & Eaton [2] point out behaviour before entering the university, motivation and commitment as relevant factors when deciding to abandon. Spady [27, 28] finds social integration, academic situation, social-economic status, gender, career and grades average for each period as relevant predictors. Other models consider the fit between the student and the institution [22, 31], institution's services and characteristics [3] and economic costs and benefits [7, 29]. More recently, Cabrera, Perez & López [8] study the problem of student retention combining psychological, social-economic and achievement mainstreams. In this last model, retention not only depends on the ability of the student to adapt, but also on the ability of the institution to adapt to the students it receives. Thus, institutions ought to implement strategies in that direction.

From another perspective, engineers have developed several models to predict student dropout from data stored by institutions, such as school grades, university admission tests, gender, economic income, among others. Díaz

978-1-7281-8328-2/20/$31.00 ©2020 IEEE

[13] uses Kaplan-Meier models and Cox proportional risks; Fischer [17] combines clustering, decision trees and neural networks. Pino [25] uses decision trees, association rules and multinomial logistic regression. Other works design a scheme to store data that may allow an early detection of possible abandoning students [15] and contextualise causes and factors for student dropout [16].

It would be valuable for the DIINF to be able to predict which students of the 6-year Informatics Engineering programme are at higher risk of dropping out. In that regard, this work intends to determine which variables allow the creation of machine learning models capable of discriminating, after completing the first semester, which students might desert because of academic factors by the end of the first year.

Early prediction of student dropout, along with the identification of key factors influencing this phenomenon, will allow an improvement in the decision-making process regarding the development of new actions and strategies to provide first-year students with academic support and assistance in their adaptation to university life. Such strategies will be designed aiming to reduce the impact of the identified factors, thus decreasing the dropout rate.

2. Data and methods

The dataset for this work considered 206 first-year students enrolled at the 6-year undergraduate programme imparted by the DIINF, cohorts 2012–2016 (excluding 2015, since it was an irregular academic year). It includes 40 datapoints for each student, including academic performance in the first semester, University Selection Test[1] results [12], social-economic and demographic data (e.g. family income quintile, city, human development index, head of household, profession of the head of household, family size, scholarships, etc.) [11]. Class variable has two possible values: Retention for the 146 students who enrolled their second year of studies, and Dropout for the 60 students who deserted as a result of academic failure. Hence, the dataset was unbalanced.

We used the random Forest algorithm (RF) in [5, 6] for the feature selection process, in order to classify students as either deserting or continuing after the end of their first year of study. RF uses a decision tree algorithm to build many simple trees with a subset of features through a bagging process. The RF parameters utilised were: number of trees $ntree = 600$, number of subset features $mtry = \sqrt{p}$, where p is the number of features initially considered for this study. The importance of each variable is calculated during the classification process, which can be measured using mean decrease accuracy (MDA) and mean decrease Gini (MDG).

The feature selection process consisted of reducing the number of variables by removing those with lowest MDA value from the set. The RF algorithm is applied repeatedly until the classification error suffers a clear increment, which could mean that an important feature was removed. Once the most important features are identified, the simplest RF model is then selected under Occam's razor criteria.

TABLE 1. CONFUSION MATRIX OBTAINED USING RANDOM FOREST WITH ALL 40 FEATURES.

Class	Retention	Dropout	Error
Retention	140	10	0.067
Dropout	6	46	0.115

TABLE 2. INITIAL RELEVANT FEATURES, IN ORDER OF IMPORTANCE.

Variable	Retention	Dropout	MDA	MDG
AI	0.080	0.213	0.113	21.782
WAG	0.054	0.216	0.095	18.820
FSE	0.017	0.080	0.033	8.121
SFR	0.016	0.072	0.030	7.974
ASR	0.007	0.053	0.018	6.148
PF	0.005	0.038	0.013	5.447
IQ	0.006	0.032	0.013	3.661

Once the relevant subset of features was identified, a decision tree was built as to obtain rules which enable knowledge extraction. For this purpose, the pruning concept was applied with a confidence factor $CF = 0.2$, which corresponds to its own generalisation process.

3. Results

3.1. Random forest and feature selection

Table 1 shows the confusion matrix obtained after applying RF to the entire dataset.

7 variables where assessed as relevant after the feature selection process, as shown in table 2, where the importance of each variable was measured using MDA and MDG. Of these, 6 are first-semester academic factors, namely weighted average grade (WAG), academic index (AI), average success rate (ASR), success-to-failure ratio (SFR), efficiency (FSE) and progress factor (PF). The remaining variable, the family income quintile (IQ) corresponds to a social-economic factor. Table 3 shows the confusion matrix obtained after applying RF using only the selected features.

3.2. Decision tree

After identifying relevant variables, we used them to build a decision tree in order to identify relationships among these features. For this purpose, we used all data for the training process, as well as a 0.2 confidence factor. Figure 1 shows the obtained decision tree and table 4 presents its associated confusion matrix.

The rules present in the model are:

TABLE 3. CONFUSION MATRIX OBTAINED USING RANDOM FOREST WITH SELECTED FEATURES ONLY.

Class	Retention	Dropout	Error
Retention	143	7	0.047
Dropout	6	46	0.115

[1] Chilean 2003–2019 standardised higher education admission test.

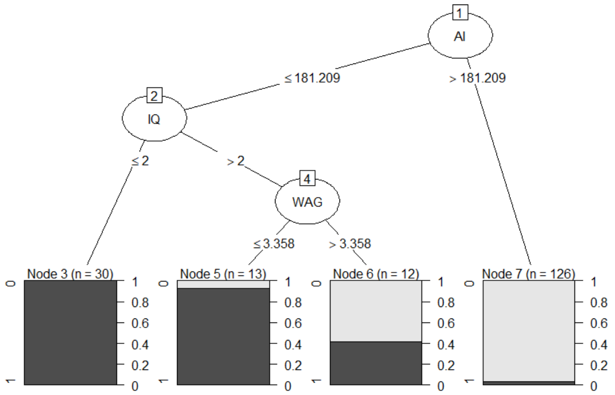

Figure 1. Obtained decision tree.

TABLE 4. Training confusion matrix for the decision tree.

Class	Retention	Dropout	Error
Retention	129	4	0.030
Dropout	1	45	0.022

- Students with an academic index above 181.209 persist.
- Students with an academic index up to 181.209 whose family income is at the two lowest quintiles abandon.
- Of the students with an academic index up to 181.209 whose family income is at the three highest quintiles, if their weighted average grade is of 3.358 or lower, they dropout. Otherwise, they persist.

When testing the model with all the instances pertaining to the 2017–2018 cohorts, results showed that the model is over-fit, as illustrated by the resulting confusion matrix (table 5).

4. Discussion and conclusion

Results of the feature selection process, shown in table 2, shed light on the variables with highest predictive value for the dropout risk for first-year students. The fact that 6 of the relevant variables (WAG, AI, ASR, SFR, FSE and PF) correspond to academic factors coincides with Tinto's conclusions [31]. Also, the remaining variable (IQ), which reflects a low family income, is consistent with the ideas proposed by Cabrera et al. [7] and St. John et al. [29]. It is interesting to note that all the selected features support the Dropout class, despite the fact that the sample is highly unbalanced.

TABLE 5. Training confusion matrix for the final decision tree.

Class	Retention	Dropout	Error
Retention	55	1	0.018
Dropout	14	9	0.609

The MDG metric is an average of the decrease in Gini that a variable produces in every tree in the forest. In a single tree, the Gini index measures the inequality within a population, so a low Gini index indicates that all the elements in a child node are more alike (i.e., belong mostly to the same class). Therefore, a high MDG for a variable means a high contribution of that variable to the model's performance. The MDA metric also assesses the importance of a variable, measuring the model's decrease in accuracy after adding noise to a variable. The selected features are equally ranked in importance by both metrics. AI holds the first place, which is highly interesting because it combines WAG, SFR and PF, thus summarising several academic performance aspects. These three variables also appear as relevant features, holding the second, fourth and sixth places respectively. WAG appearing second in importance suggests it is an important element of the AI summary variable. The importance of the remaining academic factors significantly decreases respect of AI and WAG. IQ, the only social-economic feature, appears as the least important of the selected features, suggesting that student dropout depends mainly of academic factors.

The resulting decision tree, although over-fitted, identifies AI as the main decision factor, which concurs with the importance established for this feature in the feature selection process. It must be noted that this single variable, through high AI values, detects most of the retained students. Among students with lower values of academic index, a low family income appears as the main reason for deserting, followed by a low weighted average grade. It must be noted that, in Chile, grades go from 1.0 to 7.0, where 4.0 is the minimum approving grade. Therefore, an inflection point at a WGA of 3.358 might mean that students with a rather high family income desert if they fail several courses. It is interesting to note that IQ appears higher than WAG in the tree, even though it's importance is lower according to the feature selection process. This may be explained by the fact that AI summarises WAG along with other academic variables.

It is important to highlight that results confirm the importance of factors mentioned in the main theoretical models to explain student attrition [31, 7, 27, 28], which have been proposed by authors in fields such as education, sociology and psychology. They also show that, although the mentioned models were built in a different context, they might also be valid for the Chilean reality. Therefore, machine learning techniques such as decision trees emerge as a promising means of aplying such models for early detection of at-risk students.

As a work in progress, there is still much to be done. The next step will be to build a fine-tuned decision tree, and to extend this work to the 4-year programme imparted at the DIINF. Then, we should build other models using different machine learning techniques addressed in the literature, such as support vector machines and multinomial logistic regression [9, 10, 25].

Results of the feature selection process suggest that performing cluster analysis could be a useful tool to get a better characterisation of the groups of students at higher dropout risk, by identifying other variables that may be strongly correlated to the selected ones, hence providing a broader view of these students and their situation.

Acknowledgments

This study was partially funded by the National Agency for Research and Development (ANID), Scholarship Programme, DOCTORADO BECAS CHILE, 7608/2020.

References

[1] Louis C. Attinasi. "Getting in: Mexican American students; perceptions of their college-going behavior with implications for their freshman year persistence in the university". In: *ASHE 1986 Annual Meeting Paper, San Antonio, TX.(ERIC No. ED 268 869).* 1986.

[2] John Bean and Shevawn Bogdan Eaton. "The psychology underlying successful retention practices". In: *Journal of College Student Retention: Research, Theory & Practice* 3.1 (2001), pp. 73–89.

[3] Joseph B. Berger and Jeffrey F. Milem. "Organizational Behavior at Colleges and Student Outcomes". In: *The Review of Higher Education* 23 (Dec. 2000), pp. 268–338. DOI: 10.1353/rhe.2000.0001.

[4] John M. Braxton, Anna V. Shaw Sullivan, and Robert M. Johnson. "Appraising Tinto's theory of college student departure". In: *Higher education: Handbook of theory and research* 12 (1997), pp. 107–164.

[5] Leo Breiman. "Random Forests". In: *Machine Learning* 45.1 (Oct. 2001), pp. 5–32. ISSN: 1573-0565. DOI: 10.1023/A:1010933404324. URL: https://doi.org/10.1023/A:1010933404324.

[6] Leo Breiman et al. *Package 'randomForest'*. 2018. URL: https://cran.r-project.org/web/packages/randomForest/randomForest.pdf.

[7] Alberto F. Cabrera, Nora Amaury, and María B. Castañeda. "College persistence: Structural equations modeling test of an integrated model of student retention". In: *Journal of Higher Education* 64.2 (1993), pp. 123–139.

[8] Alberto F. Cabrera, Paulina Pérez, and Lorena López. "Evolución de perspectivas en el estudio de la retención universitaria en los EEUU: bases conceptuales y puntos de inflexión". In: *Persistir con éxito en la universidad: De la investigación a la acción*. Ed. by Pilar Figuera. Barcelona, España: Laertes, 2015, pp. 15–40.

[9] Kristof Coussement et al. "Predicting student dropout in subscription-based online learning environments: The beneficial impact of the logit leaf model". In: *Decision Support Systems* 135 (2020), p. 113325.

[10] Dursun Delen. "A comparative analysis of machine learning techniques for student retention management". In: *Decision Support Systems* 49.4 (2010), pp. 498–506.

[11] Departamento de Estudios. *Registro de estudiantes de primer año de Ingeniería Civil en Informática*. Santiago, Chile, 2019.

[12] Departamento de Evaluación, Medición y Registro Educacional. *Portal Bases de Datos*. URL: https://demre.cl/portales/portal-bases-datos (visited on 08/11/2020).

[13] Christian J. Díaz. "Factores de deserción estudiantil en ingeniería: una aplicación de modelos de duración". In: *Información tecnológica* 20.5 (2009), pp. 129–145.

[14] Corinna A. Ethington. "A psychological model of student persistence". In: *Research in higher education* 31.3 (1990), pp. 279–293.

[15] Mariana Falco and Sergio Antonini. "Deserción Universitaria: Estado de Arte y Propuestas en las Regionales de la UTN, Análisis en la Regional La Plata". In: *3ºCongreso Nacional de Ingeniería Informática /Sistemas de Información (CoNaIISI)*. Universidad Tecnológica Nacional – Facultad Regional Buenos Aires, 2015.

[16] Mariana Falco, Romina Istvan, and Antonini Sergio. *University Desertion: Analysis to 2017 admission course in Information Systems Engineering*. JAIIO, 2017.

[17] Erwin Fischer. "Modelo para la automatización del proceso de determinación de riesgo de deserción en estudiantes universitarios". PhD Thesis. Santiago: Universidad de Chile, 2012.

[18] Martin Fishbein and Icek Ajzen. *Belief, Attitude, Intention and Behavior: An Introduction to Theory and Research*. Addison-Wesley, 1975.

[19] Luis Eduardo González and Daniel Uribe. "Estimaciones sobre la "repitencia" y deserción en la educación superior chilena. Consideraciones sobre sus implicaciones". In: *Calidad en la Educación* 17 (2002), pp. 75–90.

[20] Marco Kremerman and Alexander Páez. *Endeudar para gobernar y mercantilizar: El caso del CAE*. Chile, Estudios de la Fundación Sol, 2018.

[21] A. Lara, L. Elizalde, and R. Rolando. *Retención de primer año de educación superior*. Tech. rep. Santiago: MINEDUC, 2014.

[22] Oscar T. Lenning. "Variable-selection and measurement concerns". In: *New directions for institutional research* 1982.36 (1982), pp. 35–53.

[23] Ley 19.986. *Ley de presupuestos del sector público para el año 2005*. Santiago, Chile, Dec. 2004.

[24] Ley 20.027. *Establece normas para el financiamiento de estudios de educación superior*. Santiago, Chile, June 2005.

[25] Y. Pino. "Modelo de predicción de éxito académico de alumnos de la facultad de ingenieria con causales de eliminación". In: Santiago: Universidad de Santiago de Chile, 2017.

[26] Servicio de Información de Educación Superior. *Informe retención de 1er año de pregrado — Cohortes 2014 – 2018*. Oct. 2019. URL: https://www.

mifuturo.cl/wp-content/uploads/2019/10/Informe-de-Retencion_SIES_2019-octubre.pdf (visited on 05/31/2020).

[27] William G. Spady. “Dropouts from higher education: An interdisciplinary review and synthesis”. In: *Interchange* 1 (1970), pp. 64–85.

[28] William G. Spady. “Dropouts from higher education: Toward an empirical model”. In: *Interchange* 2 (1971), pp. 38–62.

[29] Edward P. St. John et al. “Economic influences on persistence reconsidered”. In: *Reworking the student departure puzzle*. Ed. by John M. Braxton. Nashville, Tennessee, USA: Vanderbilt University Press, 2000, pp. 29–47.

[30] Subdirección de Calidad y Mejora Continua. *Retención de estudiantes de las carreras de Ingeniería para el año 2019*. Santiago, Chile, 2020.

[31] Vincent Tinto. “Dropout from higher education: A theoretical synthesis of recent research”. In: *Review of Educational Research* 45.1 (1975), pp. 89–125.