Detection of Desertion Patterns in University Students Using Data Mining Techniques: A Case Study

Dayana Vila[1], Saúl Cisneros[1], Pedro Granda[1], Cosme Ortega[1],
Miguel Posso-Yépez[2], and Iván García-Santillán[1](✉)

[1] Department of Software Engineering, Faculty of Applied Sciences,
Universidad Técnica del Norte, Ibarra, Ecuador
{dpvilae, sacisnerosb, pdgranda, mc.ortega, idgarcia}@utn.edu.ec

[2] Faculty of Education, Science and Technology, Universidad Técnica del Norte,
Ibarra, Ecuador
maposso@utn.edu.ec

Abstract. Student desertion is a phenomenon that affects higher education and academic quality standards. Several causes can lead to this issue, the academic factor being a potential reason. The main objective of this research is to detect dropout patterns in the "Técnica del Norte" University (Ecuador), based on personal and academic historical data, using predictive classification techniques in data mining. The KDD (Knowledge Discovery in Databases) process was used to determine desertion patterns focused on two approaches: (i) Bayesian, and (ii) Decision Trees, both implemented on Weka. The classifiers performance was quantitatively evaluated using the confusion matrix and quality metrics. The results proved that the technique based on decision trees had slightly better performance than the Bayesian approach on the processed data.

Keywords: Student desertion · Pattern discovery · Data mining
KDD · Weka

1 Introduction

1.1 Problem Statement

Student desertion is a phenomenon that refers to the abandonment of education. This can happen due to several reasons, such as: academic, social, economic and psychological situations. Dropping out affects not only students, but also the prestige of the education center [1]. This situation is common within Ecuador's higher education, especially in the lower levels. Eight out of ten students that started public college in the year 2012 continued their studies in 2013, and seven out of ten continued in 2014 [2]. This phenomenon contributes to the waste of limited public resources designated for education, decreasing quality standards, and somehow, both rising unemployment and poverty in the country [3].

M. Botto-Tobar et al. (Eds.): CITT 2018, CCIS 895, pp. 420–429, 2019.
https://doi.org/10.1007/978-3-030-05532-5_31

Early identification of university students who have a greater probability of abandoning their career would help improve the quality indicators in higher education. Applying data mining techniques has had a significant impact in the educational sector in recent years [3, 4]. Having a predictive system able to detect students with a high probability of dropping out and, even more, providing patterns of potential deserters is important to propose and apply preventive action plans to contribute to the improvement of student academic performance and mitigate the adverse effects previously mentioned.

The student welfare department of the “Técnica del Norte” University [5] is the entity in charge of, among other things, helping university students to continue their studies. However, the students most susceptible to dropping out have been detected and reported too late. For this reason, it is necessary to implement a system that allows opportune decision making with the purpose of mitigating this phenomenon, motivating and helping the students to continue with their college studies and life plans. This phenomenon negatively affects, mainly, undergraduates in the lower semesters (up to the fourth level).

A predictive model of pattern detection is proposed in this study, which detects college dropouts in the “Técnica del Norte” University (Ecuador), using the personal and academic data of undergraduate students (bachelor’s degree and engineering) from the last 5 years. The data were processed following the KDD process (Knowledge Discovery in Databases) and using Bayesian and decision trees data mining techniques [6]. Identifying patterns considering, as possible desertion factors, the ethnic group, disability level (physical, mental, intellectual or sensory), gender, and place of residence makes up the main contribution in this research. Additionally, this study is performed using free software (Weka), fulfilling the Executive Decree of Ecuador No. 1014 [7]. This regulation disposes to use open source software within computer systems and equipment in the public administration of the Ecuadorian government. This makes up the second contribution in this study.

1.2 Review of Literature

Some existing studies were relevant for this research because they were carried out in the educational field of higher education using, among others, predictive data mining techniques. These studies are listed below:

Lehr et al. [8] used data mining techniques to predict college student retention at Embry-Riddle Aeronautical University (USA). To accomplish this goal, the data of 972 students, enrolled in 2008, were used to create predictive models, taking into account previous preparation of the students, grades in the first year, and personal and financial data. The classification techniques used were: Logistic Regression, Naive Bayes, K-Nearest Neighbors, Random Forest, Multilayer perceptron, and Decision trees. Logistic regression obtained the best results according to the error rate.

Hernández et al. [3] developed a predictive system that detects students with high desertion probabilities and the profile of dropouts in the Information and Communication Technologies Engineering career, using logistic regression, clustering, decision trees and neural networks. The results obtained by using logistic regression were best fitted to the study.

Peralta et al. [9], using data mining techniques and statistic models, identified the most relevant variables which predicted desertion and graduation rates of students in the Temuco Catholic University (Mexico). The original database held 15183 students and 143 variables. They used linear statistic models, such as the *Probit* and *Logit* models, as well as the decision tree technique. The main result obtained was that there is a clear relation among variables that depend on factors such as average and ranking.

Merchan and Duarte [10] performed a predictive model of academic performance based on the academic and demographic data of 932 students in the Computer Science career in the “El Bosque” University (Colombia). J48, PART, and Ridor decision tree algorithms were used. The J48 algorithm proved to be the best, classifying correctly around 78% of new instances.

Zaffar et al. [11] carried out research about the performance of classifiers and feature selection algorithms on a set of student data. The evaluation of the algorithms’ effectiveness was performed through precision, recall, and f-measure. The best result was obtained using the Random Forest classifier.

Devasia et al. [1] proposed a web application to extract useful information about students dropping out in higher education. The experiment was carried out using the data of 700 students containing 19 attributes. The Naïve Bayes algorithm was the most precise compared to other methods such as decision trees, neural networks, and logistic regression.

Kotsiantis [4] worked with stored data belonging to the Hellenic Open University (Greece), a distance learning program. The main objective was to predict student grades, which helped the tutors better comprehend the population characteristics.

Mishra et al. [12] used clustering tasks on the data of 84 undergraduate students of psychology in the north of Spain, grouping them into three clusters according to the course final grades. The smallest cluster was the most understood while the two bigger clusters were more difficult to interpret.

Moscoso-Zea et al. [13] designed a data warehouse in an educational scenario, using the methodology proposed by Kimball and Ross [14]. This repository is intended to the knowledge discovery process and indicators analysis of academic performance in future researches.

Next, the research methodology used in this study is presented.

2 Materials and Methods

2.1 The KDD Process

The KDD (Knowledge Discovery in Databases) process was used to determine student desertion patterns and it consists of four phases [6]: (i) data collection; (ii) selection, cleansing and transformation of data; (iii) data mining; and (iv) evaluation and interpretation of the model. This research used personal and academic students data from 37 undergraduate careers of the UTN (25 bachelor’s degrees and 12 engineering) from last five years (2013–2017). Each phase of the KDD process used in this study is detailed below.

2.2 Data Collection

The purpose in this stage is to integrate all data into a single repository (data warehouse). The main source of the raw data was the academic database of the UTN, which is stored on an Oracle 11g server, containing 17.882 student records. The data are both qualitative and quantitative in nature.

A part of the database was selected, i.e. 9 tables containing personal and academic data. Thirteen variables were considered in this study: age, ethnicity, nationality, place of origin, place of residence, disability level, gender, marital status, family burdens, study modality, faculty, average grades, and current state. The current state variable (active/inactive; YES/NO) was the class attribute, that is, the qualitative variable used to predict the classification (student desertion). The variables age, average grades and disability level are originally of a quantitative nature, while the rest are qualitative attributes.

Some variables mentioned above are relevant for this research, considering that the UTN is in Zone-1 of Ecuador (made up of the provinces: Imbabura, Carchi, Esmeraldas, and Sucumbíos). This zone has 638.979 men and 634.353 women of the following ethnicities: 60.9% mestizos, 21.7% Afro-Ecuadorians and African descendants, 11.5% indigenous, and 4.3% Caucasian white [15]. Also, 36.878 people (2.9%) of this region have a disability [16]. This research is based on the pattern identification in student desertion considering such variables (ethnicity, disability, gender, and place of residence) as possible dropout factors.

A data warehouse was built from the raw data, using PostgreSQL version 9.4 [17]. To accomplish this step, the data were exported from Oracle to a MS Excel file in *xlsx* format, and then imported into PostgreSQL. This database management system was selected due to it is open source and also its easy connection to the Pentaho Community Suite [18], which is used in the next phase.

2.3 Selection, Cleansing, and Transformation of Data

This phase's goal is to obtain clean data, without null or inconsistent values, allowing a highly reliable student desertion patterns. The data selection, cleansing, and transformation tasks were performed using the Pentaho Community Suite, obtaining a mineable view, that is, a dataset ready for applying data mining techniques [13].

This study used only data from the first four levels, considering that the student desertion phenomenon in the UTN frequently occurs at lower levels. The data transformation converted discrete quantitative variables (age, average grade, and disability level) into ordinal qualitative attributes using the categories indicated in Table 1. The age attribute was divided into three categories: low, medium and high. The low level was considered until the 25 years of age because the undergraduate students finish their career mostly in this period. The medium level up to 40 years because the postgraduate students (master's degree) are common in this age range. The high level is considered more than 40 years because here the students are less frequent. The average grade variable was divided into three categories (low, medium and high) considering the academic regulation in the UTN. The disability level was divided into four categories (slight, moderate, severe and very severe) considering the disability regulation of Ecuador [19].

After cleansing and transforming the data, 12620 instances were obtained, which conform the mineable view. Next, it was exported to a plain text file delimited by commas (*.*csv*) for the following data mining phase.

Table 1. Categories used in the transformation of quantitative data

Attribute	Category	Interval
Age	Low	≤ 25 years
	Medium	26–40 years
	High	≥ 41 years
Average grade	Low	≤ 7 points
	Medium	8–9 points
	High	9–10 points
Disability level	Slight	30–49%
	Moderate	50–74%
	Severe	75–84%
	Very severe	85–100%

2.4 Data Mining

This phase aims to obtain predictive models representative of the data, using the WEKA software version 3.8 [20]. This software is open source and it contains a wide collection of machine learning algorithms to generate different models.

In this study, the prediction used classification techniques, which consists in finding a pattern in an uncategorized data group and classify it into a predefined set of classes [6]. The classifiers used were the following: (i) decision trees with the RandomTree algorithm, due to the simplicity of its model, easy interpretation, and speed to classify new data [6, 21, 22]; and (ii) Bayesian technique with the Naïve Bayes algorithm, because of its simplicity and high precision in several domains [1, 21].

2.5 Evaluation and Interpretation of the Predictive Model

In this stage, the predictive models' performances were evaluated quantitatively. Subsequently, the best model was used to obtain student desertion patterns. The Cross-validation method (10-fold) was used for evaluating the models. For each classification algorithm, the following quality metrics were used [21, 22]: accuracy, error rate, Kappa coefficient, TP rate, FT rate, precision, recall, F1 score, and ROC area. The interpretation of the best model was carried out with the help of an expert using visualization techniques, which facilitate comprehension and discovery of desertion patterns.

Next, the results obtained in this study are presented.

3 Results and Discussion

3.1 Evaluation of Predictive Models

The overall evaluation of the classification was performed by analyzing a confusion matrix and computing several quality metrics [23], which are shown in Table 2.

Table 2. Overall evaluation metrics for the classifiers used in this study

Classifier	Accuracy	Error rate	Kappa coefficient
RandomTree	97.607%	2.393%	0.2046
Naive Bayes	97.544%	2.456%	0.2189

The classifier with the greatest accuracy and the lowest error rate turned out to be the RandomTree algorithm. Regarding the Kappa coefficient, both classifiers keep a "Fair" strength of agreement, according to the classification proposed by Landis and Kock [24]. The coefficient Kappa indicates whether the results obtained in the confusion matrix are significantly better than those produced in a random classification [25].

Some specific metrics used in the evaluation of the classifiers are shown in Table 3. It is noted that similar values were obtained for both classifiers, except for the ROC area, where it was lower for the RandomTree. This discrepancy RandomTree having greater accuracy but a lower ROC area than Naïve Bayes, may be due to an imbalanced dataset [26, 27], i.e. a disparity in the frequencies of the observed classes. This is detailed later.

Table 3. Specific evaluation metrics for the classifiers used in this study (weighted average).

Classifier	TP rate	FP rate	Precision	Recall	F_1 score	ROC area
RandomTree	0.976	0.839	0.968	0.976	0.970	0.836
Naive Bayes	0.975	0.822	0.967	0.975	0.970	0.925

Table 4 shows the confusion matrices generated by both RandomTree and Naive Bayes algorithms, respectively.

Table 4. Confusion matrices for classification algorithms.

Classifier	TP	FN	FP	TN	Number of instances
RandomTree	41	249	53	12277	12620
Naive Bayes	46	244	66	12264	12620

For the RandomTree classifier, 12318 instances were correctly classified and 3022 incorrectly. The results of the errors are the following:

- This classifier attempted to classify 290 examples of the class status = YES. Of these, 41 instances were correctly classified and 249 incorrectly classified as class status = NO.
- It tried to classify 12330 examples of the class status = NO. Of these, 12277 instances were correctly classified and 53 incorrectly classified as class status = YES.

For the Naive Bayes classifier, 12310 instances were correctly classified and 310 incorrectly. The results of the errors are the following:

- This classifier attempted to classify 290 examples of the class status = YES. Of these, 46 were correctly classified and 244 incorrectly classified as class status = NO.
- It tried to classify 12330 examples of the class status = NO. Of these, 12264 were correctly classified and 66 incorrectly classified as class status = YES.

Notably within the confusion matrices of both classifiers (Table 4), the cases for the number of negative examples NO (12330, 97.7%) is much higher than the positive examples YES (290, 2.3%), i.e. it is dealing with imbalanced dataset as mentioned before, than most supervised learning methods will skew the predicted probabilities, tending to predict the abundant class more often. In this case, the goodness of a classifier is best approximated using the F_1 score metric, which can be interpreted as a weighted average of the precision and recall values [21, 28].

In this way, considering F_1 score metric (Table 3), both classifiers keep similar performances. However, RandomTree provides a set of decision rules (representation equivalent to the decision tree) that facilitates the processing and interpretation of the predictive model [3]. Therefore, it was chosen in this study for pattern discovery in students' desertion. The decision rules generated by this model can be seen online in Appendix 1 from https://bit.ly/2Mmy9a7. In contrast, the Naïve Bayes algorithm does not generate a model but rather it performs the classification just at the requested time.

3.2 Detection of Student Desertion Patterns

After selecting the best predictive model (tree and decision rules), the exploration and interpretation of the tree was completed from the root toward the leaves (class attribute) labeled with status = YES [22]. This led to the following student desertion patterns being identified:

a. Older married women studying arts-related careers with a low average.
b. Older married men residing in Imbabura and maintaining a low average.
c. Younger married mestizo people who study in blended modality and live outside Imbabura.
d. Single, indigenous people, regardless of their sex, do not reside in Imbabura and keep a medium or low average.
e. Women in free union who study in the classroom-based modality with a medium or low average.
f. People of Colombian nationality.
g. Women with moderate disabilities and medicine-related studies.
h. People in free union with studies related to health, art or agricultural sciences.

i. Afro-Ecuadorian people studying engineering.
j. Divorced men studying engineering.

The patterns of student desertion identified above are very useful for the UTN's student welfare department and these will allow to propose action plans to mitigate such phenomenon.

3.3 Discussion

The desertion patterns identified in this study coincide with some results from prior researches. Merchan and Duarte [10] indicated that (i) students in free union are potential deserters, whereas in this work there is a higher desertion correlation with students studying health, art or agricultural sciences. Some commonalities with other works that were used in the discovery of student desertion patterns are the place of origin, average grade, study modality [1, 3], whereas some variables that are not consider in this study are the cumulate average and the ranking [9]. In addition, this study considers the data of students from 37 undergraduate careers of the UTN, unlike other works where a specific career is considered [3, 10].

Some limitations in this study are the following: (i) the problem of detection of student desertion patterns is addressed only from the academic point of view, i.e. based on average grades; and (ii) predictive classification techniques are solely used for this purpose.

Finally, some future works hold this research line are the following: (i) expand the study also considering socioeconomic and psychological data; (ii) generate other predictive models based on logistic regression and neural networks; (iii) include descriptive tasks of data mining as the clustering [3, 12], association, outlier detection; (iv) create and compare specific models for each faculty of the UTN.

4 Conclusions

Detection of student desertion patterns in the university academic field is addressed in this study by using predictive classification tasks based on the decision tree (RandomTree) and Bayesian (Naïve Bayes) approaches. The student data used are personal and academic types from 37 undergraduate careers of the "Técnica del Norte" University (Ecuador), and these are processed following the KDD process (Knowledge Discovery in Databases) and using free software (Weka). The classifiers evaluation is quantitatively analyzed considering quality metrics. The RandomTree algorithm provides, slightly, a better performance than Naive Bayes (Table 2). The interpretation of the best model allows to identify relevant patterns of student desertion (Sect. 3.2), which contributes to propose action plans to mitigate this phenomenon. As future work, it is recommended to consider socio-economic and psychological data and include descriptive tasks of data mining.

Acknowledgment. To the IT department of the "Técnica del Norte" University for allowing access to the raw data of the academic database.

References

1. Devasia, T., Vinushree, T.P., Hegde, V.: Prediction of students performance using Educational Data Mining. In: International Conference on Data Mining and Advanced Computing (SAPIENCE), Ernakulam, pp. 91–95 (2016). https://doi.org/10.1109/sapience.2016.7684167
2. Senescyt: Rendición de cuantas 2015 (Accountability 2015). Quito-Ecuador (2015). http://www.senescyt.gob.ec/rendicion2015/assets/presentacion-rendicion-de-cuentas.pdf. Accessed 13 Mar 2018
3. Hernández, G., Melendez, R.A., Morales, L.A., Garcia, A., Tecpanecatl, J.L., Algredo, I.: Comparative study of algorithms to predict the desertion in the students at the ITSM-Mexico. IEEE Latin Am. Trans. **14**(11), 4573–4578 (2016). https://doi.org/10.1109/tla.2016.7795831
4. Kotsiantis, S.B.: Use of machine learning techniques for educational proposes: a decision support system for forecasting students' grades. Artif. Intell. Rev. **37**(4), 331–344 (2012). https://doi.org/10.1007/s10462-011-9234-x
5. UTN: Universidad Técnica del Norte (2018). http://www.utn.edu.ec/. Accessed 22 Apr 2018
6. Lara, J.: Minería de Datos (Data Mining). Madrid, Centro de Estudios Financieros (2014)
7. Subsecretaría de Informática: Decreto Ejecutivo No. 1014 (Executive Decree No. 1014) (2009). http://cti.gobiernoelectronico.gob.ec/ayuda/manual/decreto_1014.pdf. Accessed 22 Apr 2018
8. Lehr, S., Liu, H., Kinglesmith, S., Konyha, A., Robaszewska, N., Medinilla, J.: Use educational data mining to predict undergraduate retention. In: IEEE 16th International Conference on Advanced Learning Technologies (ICALT), Austin, TX, pp. 428–430 (2016). https://doi.org/10.1109/icalt.2016.138
9. Peralta, B., Poblete, T., Caro, L.: Automatic feature selection for desertion and graduation prediction: a chilean case. In: 35th International Conference of the Chilean Computer Science Society (SCCC), Valparaiso, pp. 1–8 (2016). https://doi.org/10.1109/sccc.2016.7836055
10. Merchan, S.M., Duarte, J.A.: Analysis of data mining techniques for constructing a predictive model for academic performance. IEEE Latin Am. Trans. **14**(6), 2783–2788 (2016). https://doi.org/10.1109/TLA.2016.7555255
11. Zaffar, M., Hashmani, M.A., Savita, K.S.: Performance analysis of feature selection algorithm for educational data mining. In: IEEE Conference on Big Data and Analytics (ICBDA), Kuching, pp. 7–12 (2017). https://doi.org/10.1109/icbdaa.2017.8284099
12. Mishra, A., Bansal, R., Singh, S.N.: Educational data mining and learning analysis. In: 7th International Conference on Cloud Computing, Data Science and Engineering - Confluence, Noida, pp. 491–494 (2017). https://doi.org/10.1109/confluence.2017.7943201
13. Moscoso-Zea, O., Andres-Sampedro, Luján-Mora, S.: Datawarehouse design for educational data mining. In: 15th International Conference on Information Technology Based Higher Education and Training (ITHET), Istanbul, pp. 1–6 (2016). https://doi.org/10.1109/ithet.2016.7760754
14. Kimball, R., Ross, M.: The Data Warehouse Toolkit. Wiley, Indianapolis (2013)
15. INEC: Ecuador en Cifras (Ecuador in figures) (2010). http://www.ecuadorencifras.gob.ec/resultados/. Accessed 10 Nov 2018
16. Conadis: Estadística y análisis de datos de personas con discapacidad (Statistics and data analysis of people with disabilities) (2018). http://www.consejodiscapacidades.gob.ec/wp-content/uploads/downloads/2018/03/index.html. Accessed 04 Nov 2018

17. The PostgreSQL Global Development Group: Download PostgreSQL (2018). https://www.postgresql.org/download/. Accessed 04 Sept 2018
18. HITACHI: Hitachi Vantara (2018). http://www.pentaho.com/pentaho-community-edition-5-0-now-available. Accessed 04 May 2018
19. Conadis: Reglamento a la Ley Orgánica de Discapacidades del Ecuador (Regulation to the Organic Law on Disability of Ecuador) (2017). https://www.consejodiscapacidades.gob.ec/wp-content/plugins/download-monitor/download.php?id=19&force=1. Accessed 12 Feb 2017
20. The University of Waikato: Weka 3: Data Mining Software in Java (2018). https://www.cs.waikato.ac.nz/ml/weka/. Accessed 22 Mar 2018
21. Sierra, B.: Aprendizaje automático: conceptos básicos y avanzados (Machine learning: basic and advanced concepts). Prentice-Hall, Madrid (2006)
22. Pajares, G., de la Cruz, J.: Aprendizaje automático: un enfoque práctico (Machine learning: a practical approach). Madrid, Ra-Ma (2010)
23. Castillejo-González, I.L., López-Granados, F., García-Ferrer, A., Peña-Barragán, J.M., Jurado-Expósito, M., Sánchez De La Orden, M., et al.: Object and pixel-based classification for mapping crops and their agri-environmental associated measures in QuickBird images. Comput. Electron. Agric. **68**, 207–215 (2009)
24. Landis, J.R., Kock, G.G.: The measurement of observer agreement for categorical data. Biometrics **33**, 159–174 (1977)
25. Congalton, R.G.: A review of assessing the accuracy of classification of remotely sensed data. Remote Sens. Environ. **37**, 35–46 (1991)
26. Saito, T., Rehmsmeier, M.: The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE **10**(3), e0118432 (2015). https://doi.org/10.1371/journal.pone.0118432. Ed. by Brock, G.
27. MathWorks (2018). Mastering Machine Learning. https://es.mathworks.com/campaigns/products/offer/mastering-machine-learning-with-matlab.html
28. Jeni, L.A., Cohn, J.F., De La Torre, F.: Facing imbalanced data–recommendations for the use of performance metrics. In: Humaine Association Conference on Affective Computing and Intelligent Interaction, Geneva, pp. 245–251 (2013). https://doi.org/10.1109/acii.2013.47