ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645
Indian Journal of Science and Technology, *Vol 12(4), DOI: 10.17485/ijst/2019/v12i4/139729, January 2019*

Predicting University Dropout through Data Mining: A Systematic Literature

Mayra Alban[1]* and David Mauricio[2]

[1]Technical University of Cotopaxi, Faculty of Computer Science and Computer Systems, Ecuador; mayra.alban@utc.edu.ec
[2]National University of San Marcos, Artificial Intelligence Group, Perú; dmauricios@unmsm.edu.pe

Abstract

Objectives: To make a systematic review of literature on the prediction of university student dropout through data mining techniques. **Methods/Analysis:** The study was developed as a systematic review of the literature of empirical research results regarding the prediction of university dropout. In this phase, the review protocol, the selection requirements for potential studies and the method for analyzing the content of the selected studies were provided. The classification presented in section 3 allowed answering the main research question. What are the aspects considered in the prediction of university student desertion through data mining? **Findings:** University dropout is a problem which affects universities around the world, with consequences such as reduced enrolment, reduced revenue for the university, and financial losses for the State which funds the studies, and also constitutes a social problem for students, their families, and society in general. Hence the importance of predicting university dropout, that is to say identify dropout students in advance, in order to design strategies to tackle this problem. **Novelty /Improvement:** This is the first work to perform an integral systematic literature review about university dropout prediction through data mining, with studies from 2006–2018.

Keywords: Data Mining, Dropout Factors, Dropout Prediction, Machine Learning, University Student Dropout

1. Introduction

There is currently an increasing interest in researching the topic of university dropout around the world[1], with one of the main concerns being elevated rates of occurrence[2]. Dropout negatively affects institutions in the reduction of enrolment and the non-achievement of institutional objectives[3]. As a consequence, students, universities and governments are affected in both economic and social terms. Furthermore, dropout becomes a critical topic when university administrators do not possess the tools necessary to identify students who are at risk of leaving the institution. In turn, potential corrective measures are reduced[4], which might have enabled student retention at higher education institutions[5]. In the same way, the early prediction of student dropout has become a major challenge, as well as identifying the factors which contribute to this increasingly occurring phenomenon[6]. One possible reason that there are still high university dropout rates may be associated with the fact that most of the prediction models applied to solve this problem are difficult to interpret[7]. A significant effort has been made to close the university dropout gap and thus reduce dropout rates. Nonetheless, this effort has been insufficient[4]; according to the Organization for Economic Cooperation and Development (OECD), in 2016, European dropout rates ranged between 30% and 50%, while in the United States the student dropout rate was 37%[8].

In some Latin American countries, such as Columbia, dropout rates exceeded 40%, while in Brazil they reached approximately 54%. In Costa Rica, the dropout rate reached 50%[9], with public universities presenting higher dropout rates than private ones[10]. One of the measures to deal with university dropout is based on predicting its rates; therefore, data mining is used, aimed at developing methods to identify patterns among large datasets and

*Author for correspondence

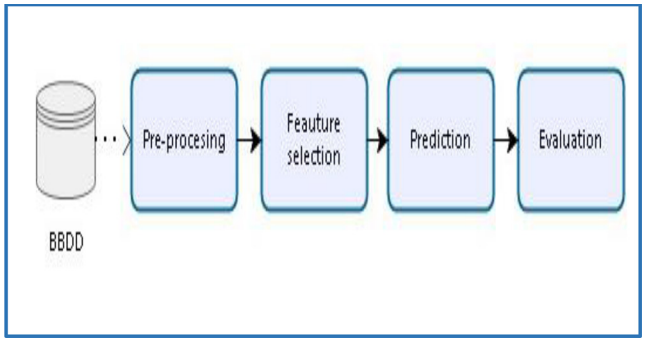

Figure 1. Data mining process for university dropout prediction[24].

thereby extract meaningful knowledge[2]. This approach is widely used in the prediction process to study dropping out, due to its acceptable degree of significance[11,12]. In general, this process follows four stages, which range from data pre-processing to result evaluation (Figure 1).

Prior literature survey on data mining and education[13,14] have covered topics such as: learning management systems, intelligent tutoring systems, adaptive educational systems, learning analytics, student modeling, and predicting academic performance. However, none of these considers the topic of university dropout, despite the large number of studies regarding factors that influence university dropout and techniques for dropout prediction. For this reason, the present study aims to answer the following question: What aspects are considered in predicting university student dropout through data mining? To meet this objective, we propose a systematic literature review of the period 2006–2018, including journals indexed in Scimago Journal & Country Rank, from which we identified and analyzed 67 articles from nine academic publishers. The present article is organized in five sections. The first section is this introduction, followed by the methodology for the systematic literature review. Subsequently, the results and analysis of the selected documents are presented in the third section. The discussion and conclusions are then presented in the fourth and fifth sections, respectively.

2. Research Methodology

In order to perform this systematic review, we considered the methodologies applied by[15], which consist in three stages:

Planning: This stage identifies the need for research and the determination of a review protocol.

Implementation: This stage implements the plan; the defined protocol is applied as well as the inclusion and exclusion criteria.

Results: This stage presents the results and statistical analysis of the selected documents.

2.1 Planning

Five research questions were proposed in order to determine the aspects that have been developed to predict university student dropout.

- Question 1 (Q1): What techniques are used for data pre-processing?
- Question 2 (Q2): What factors affect dropout?
- Question 3 (Q3): What techniques are used for factor selection?
- Question 4 (Q4): What techniques are used for prediction and what are their levels of reliability?
- Question 5 (Q5): What tools are used?

Articles from conferences and journals indexed in Scimago Journal Country Rank (SJR) with impact factor were reviewed in the following databases: Science Direct, ACM Digital Library, IEEE Xplore, Springer, DOAJ, Taylor and Francis, Emerald, Proquest and Ebsco. For document selection, the inclusion and exclusion criteria presented in Table 1 were applied.

The following search criteria were considered: "dropout student" OR "drop out student" OR "dropping student" AND "data mining", which were applied to the title, abstract and keywords in the search period between January 2006 and December 2017.

Table 1. Criteria for document selection

Inclusion	Exclusion
Models to provide a solution to the problem of university student dropout. Documents that present factors influencing university dropout. Papers that include prediction based on data mining. Papers that present metrics to assess the quality of predictive models. Papers that respond to the research questions.	Prediction documents that are unrelated to university student dropout, such as primary, secondary and postgraduate education. Documents not related to data mining. Documents that do not have numeric experimentation. Documents that are not found within the established search period.

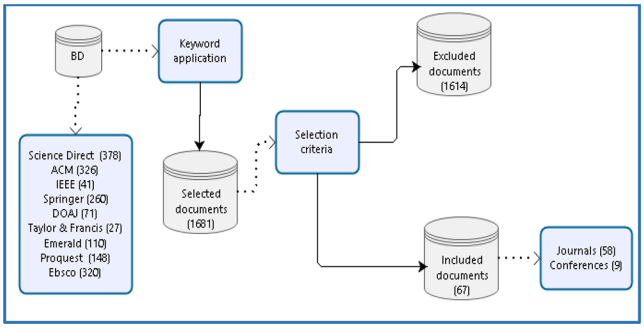

Figure 2. Systematic literature review process.

Table 2. Selected papers

Source	Identified papers	Selected papers
Science Direct	378	27
ACM Digital Library	326	1
IEEE Xplore	41	10
Springer	260	6
DOAJ	71	5
Taylor and Francis	27	5
Emerald	110	3
Proquest	148	4
Ebsco	320	6
Total	**1681**	**67**

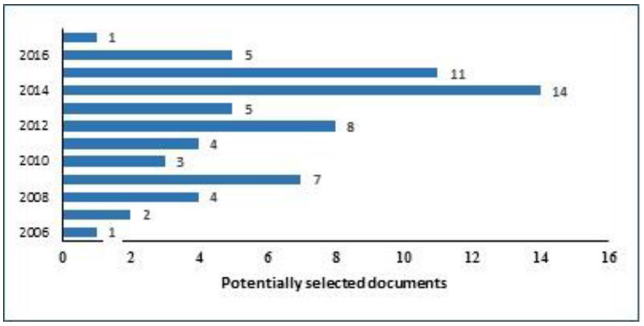

Figure 3. Temporal trend of selected publications on university dropout.

2.2 Implementation

We performed the search process based on the strategies proposed in section 2. Once selected, each document's content was reviewed in order to determine whether it matched the established selection criteria. The systematic literature review process is presented in Figure 2.

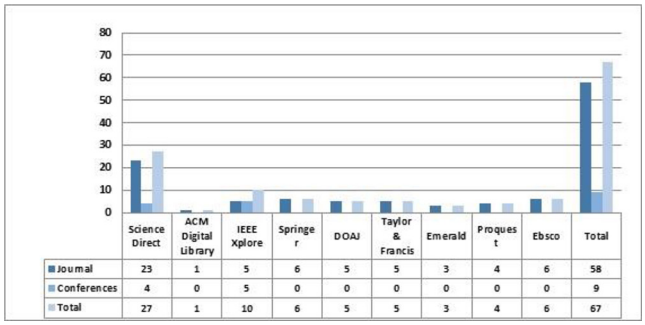

Figure 4. Publications on university dropout prediction.

3. Result

Table 2 summarizes the total identified and selected documents by information source, Science Direct being the main source of information, with 40% of the primary selected studies. Meanwhile, Emerald and ACM Digital Library present rates of 4.47% and 1.49%, respectively. Figure 3 exhibits the increase in studies during the past 12 years and the interest that researchers have in solving the problem of university dropout prediction. 87% of the primary selected documents come from journals (58 studies out of 67), and 13% correspond to publications in conferences (9 studies of 67), as presented in Figure 4. From the selected documents, we identified three aspects regarding university dropout prediction: factors, techniques and tools, all of which are specified in the framework of the present study.

Dropout factors: The reasons for which students leave studies[16].

Data mining techniques: The objective of these techniques is to discover patterns, profiles and trends through

Table 3. Techniques for data pre-processing

ID	Technique
TDP1	Multivariate analysis of variance[35]
TDP2	Bagging[36]
TDP3	Discretization[7,20,37–39]
TDP4	Attribute-based filtering[40,55]
TDP5	Single imputation[41]
TDP6	Multiple imputation[47,48]
TDP7	Normalization[19,37,26,43,42,12]
TDP8	Oversampling[22]
TDP9	Simple random sampling[43,12]

Table 4. Personal dimensionfactors

ID	Factors
PDF01	Adjustment[31]
PDF02	Age[2,7,20,25,28,44–55]
PDF03	Change of goal[28,31,56]
PDF04	Choice to change current course[12]
PDF05	Country or city of origin[40,57,29]
PDF06	Dependents[7]
PDF07	Disability[7]
PDF08	Domicile[7,44,75,31,20]
PDF09	Encouragement and support from parents[25]
PDF10	Engagement of student[28,42,56,58]
PDF11	Ethnicity[7,59,20,25,33,12,67,68,70,71,60]
PDF12	Gender[2,7,12,22,20,28,29,44,49,50,59,33,60,62-68,70,81]
PDF13	Has a computer[68]
PDF14	Health problem[28]
PDF15	Interest level in the current course[12]
PDF16	Intrinsic motivation[84,79]
PDF17	Leadership[58]
PDF18	Level of commitment[80]
PDF19	Living on campus[33]
PDF20	Loneliness[57]
PDF21	Marital status[5,22,49,68,29,73]
PDF22	Measure of student persistence[45,69]
PDF23	Pessimism[65]
PDF24	Residency[20,66,33]
PDF25	Self-efficacy[66,58,69,2]
PDF26	Student satisfaction[28,44,52,57]
PDF27	Tuition fee source[20]
PDF28	Vocational involvement[72]
PDF29	Work experience[41]
PDF30	Year of birth[36]

Table 5. Academic dimensionfactors

ID	Factors
ADF01	Absenteeism[58]
ADF02	Academic ability[75,61]
ADF03	Academic overload[75]
ADF04	Academic performance[5,61,81]
ADF05	Age at admission[65]
ADF06	Average formative assessment result[20]
ADF07	Best test score GPA[3,5,63,69,34,37,44,46,47,48,56,57,59,33,12,67,70,78,80]
ADF08	Cohort[7,60,70,71]
ADF09	Curricular involvement[68]
ADF10	Degree[5,7,36,48,49,59]
ADF11	Degree aspiration[72-78]
ADF12	Degree program length[74]
ADF13	Drop out intention[57]
ADF14	Educational goal[28]
ADF15	English language literacy[41]
ADF16	Enrolled in other institution[28]
ADF17	Entry qualifications[7,69]
ADF18	Experience[2,7]
ADF19	Final examination test[20,77,26,46,58,65]
ADF20	First and second mid-term exam grade[65]
ADF21	First semester credit load[33]
ADF22	Motive for choice[28]
ADF23	Number quiz[65]
ADF24	Participate in extra curriculum activity[28,31]
ADF25	Points from secondary[49,58,12]
ADF26	Progression outcome[7]
ADF27	Readiness[2]
ADF28	Recognized credits[61,20,29]
ADF29	Resources use[72]
ADF30	Satisfaction with course[31]
ADF31	Score of academic integration[48,59,65]
ADF32	Scores[25,38,40,29,65,79-84]
ADF33	Self-evaluation[58,68]
ADF34	Student enrolment status[58,74,12]
ADF35	Study center[20,25,68,73]
ADF36	Study level[20,41,50,33]
ADF37	Study shrift[68]
ADF38	Success rate[5,20]
ADF39	Support for learning[79]
ADF40	Total failed courses[20]

data analysis using pattern recognition technologies and advanced data analysis techniques.

Data mining tools: This refers to software used to extract patterns, trends and regularities to discover and better understand the data and predict future behavior[17].

a) Q1: What techniques are used for data pre-processing?

In the pre-processing stage, eleven techniques were identified (Table 3). This stage allows the management of anomalies as well as the correction of atypical and

missing values[17]. The purpose of these techniques is to improve the properties of the variables and solve data anomalies to optimize the search process of data mining algorithms[18]. This is based on three activities: integration, cleaning and transformation of the information. All of the studies[10] involving the pre-processing stage are concentrated around the activity of data transformation, with the techniques of normalization and discretization being the most commonly used. However, integration and cleaning activities are also important; as in[19,20] indicate; selecting the wrong variables in the data mining process can negatively affect prediction accuracy for these techniques.

b) Q1: What factors affect dropout?

We identified 112 factors to predict university dropout, which were classified according to the five dimensions (personal, academic, economic, social and institutional) proposed by author[21].

Personal factors: These constitute characteristics that determine student behavior such as feelings, thoughts or actions, which are decisive in the development of their educational environment. We identified 31 factors in the personal category, and these corresponded to approximately 28% of the total identified factors, as shown in Table 4. For many authors, personal factors are the main cause of students dropping out of university, and Table 4 evidences this fact. Age and gender are the most frequently used factors for prediction; this is because they are considered internal factors of variability which are simple to define and measure[22].

Academic factors: These refer to the development of students in their formative process. We identified 40 academic factors, which correspond to 36% of the total identified factors, presented in Table 5.

Analysis of these factors shows that the university entrance test is the most frequently used factor in the literature. However, it bears mentioning that the learning process at university has a close relationship with preceding study levels, impacting further educational achievements[23]. In the same way, the score that a student obtains in the university entrance examination is considered an indicator to explain success or failure in academic trajectory at university[5]. In this sense, many studies have analyzed the predictive validity of this factor, considering it a predictor of cognitive and attitudinal characteristics that is of the utmost importance for students to succeed at university[24].

Economic factors: These are related to students' ability to satisfy the economic requirements that present themselves during the academic program. In this dimension, 15 factors were identified that affect dropout, and they correspond to approximately 13% of the total analyzed factors, which are presented in Table 6. These economic dimension factors refer to material comforts and the ability of parents to allocate more and better resources for the academic performance of their children, which has a significant impact on academic achievements[25].

Social factors: These are aspects that affect students as a whole, and which are determined by their place and space, as presented in Table 7.

On the other hand, the social dimension focuses on the importance of the interaction between students and their social environment; interaction in relation to the institution, academic norms, and study habits[26].

Institutional factors: The factors that correspond to this category relate to the structural and functional characteristics of an institution, which are presented in Table 8; these represent approximately 3.53% of the total analyzed factors.

c) Q3: What techniques are used for factor selection?

We identified ten techniques for factor selection, which are presented in Table 9. The objective of these techniques is to select the most relevant factors used as input

Table 6. Economic dimensionfactors

ID	Factor
EDF01	Awarded scholarship[3,63,40]
EDF02	Below poverty line[22]
EDF03	Campus employment[33]
EDF04	Dependency[25]
EDF05	Fall Student Loan[63]
EDF06	Family income[68]
EDF07	Parent occupations[5,64]
EDF08	Financial concern[78,79]
EDF09	Financial need[3]
EDF10	Investment[80]
EDF11	Joint gross income of guardians[12]
EDF12	Loan received[3,63,48]
EDF13	Student employment status[58]
EDF14	Student fees status[74]
EDF15	Type of financial assistance[63,38,48,12]

Table 7. Social dimension factors

ID	Factors
SDF01	Campus accommodation[25,74,78]
SDF02	Category (marginalized or vulnerable section of society)[22,71]
SDF03	College status[44]
SDF04	Community support[58]
SDF05	Employment status[22,36,12,68]
SDF06	Family problems[31]
SDF07	Family type[31]
SDF08	Father's educational level[5,49,28,59,68,29]
SDF09	Housing indicator[49]
SDF10	Level of involvement in social media[12]
SDF11	Means of transport[68]
SDF12	Migrated before[28,60]
SDF13	Mother migrated[60]
SDF14	Mother's educational level[5,28,49,58,59,68,81]
SDF15	Occupation[29,70]
SDF16	Parent occupation[28]
SDF17	Political status[20]
SDF18	Social status[25,49,50,29,72,74]
SDF19	Stress[28]
SDF20	Student use of drugs[25]
SDF21	Use of recreational facilities[33]

variables for dropout prediction models. Approximately 55% (23 out of 42 studies) used descriptive statistics, as this technique produces the characteristics of dispersion, location and distribution of the variables[27]. Additionally, the technique is frequently used to identify patterns regarding student characteristics and behaviors related to dropout. Of these 23 studies, 14 are oriented towards variable correlation and apply this type of analysis to evaluate the association and relationship of quantitative data in terms of directionality, through correlation coefficients[28]. On the other hand, 12% (5 out of 42 studies) apply Principal Components Analysis to reduce the dimensionality of the observed variables to a number of hypothetical variables; thus, groups of variables that correlate with one another are created. These variables are transformed into independent factors that are implemented in dropout prediction models[29].

d) Q4: What techniques are used for prediction and what are their levels of reliability?

Table 8. Institutional dimensionfactors

ID	Factors
IDF1	Campus environment[31]
IDF2	High school type[67]
IDF3	Institutional involvement[72]
IDF4	Universityinfrastructure[31]

Table 9. Techniques for the selection of factors

ID	Techniques
TSF01	Analysis of variance[22,77,56,44,76]
TSF02	Descriptive Statistics[5,22,25,69,77,38,24,46,48,50,52,57,59,33,27,62,68,70–76,78–82,31,84]
TSF03	Feature extraction algorithm[39]
TSF04	Genetic Algorithm[20]
TSF05	Hosmer and Lemeshow[5,69]
TSF06	Locality Preserving Projection[26]
TSF07	Maximum Likelihood[47]
TSF08	Neighborhood Preserving Embedding[26]
TSF09	Principal Components Analysis[28,26,36,37]
TSF10	Kaiser Meyer Olkin[5,83]
TSF11	U Mann Whitney[72]

We identified14 data mining techniques, which had been classified into artificial intelligence and statistical method techniques; these are presented in Tables 10 and 11. Approximately 79% (22 out of 28 studies) used Decision tree classifiers. According in[22,30] this technique is used due to its flexibility when processing data of a numerical and categorical nature, its monotonous transformations of explanatory variables, and the ease of interpreting results. Furthermore, it presents better accuracy rates. In[31] and[32] mention that the algorithm ID3 (Decision tree classifier) is effective in classifying data from student history registers and is more sensitive in comparison to other algorithms.

Neural network classifiers and support vector machines hold the second highest frequency of use, since these data mining approaches are considered powerful tools for solving classification problems[33] and are used frequently for their simplicity and ease of understanding[32]. Four statistical techniques were identified, corresponding to a total of 36 references, or 3% (4 out of 14 techniques) of the total studies analyzed. Of these, 54% (21 out of 39 studies) applied Linear Regression and Logistic Regression, as

Table 10. Artificial intelligence techniques

ID	Technique
AI1	Neural network classifier[2,4,19,63,61,20,26,37,40,41,49,51,58,68]
AI2	Support vector machine[2,19,63,61,34,39,41,51,54,12,32]
AI3	Decision tree classifier[2,4,63,22,61,20,34,26,36,37,39,40,49,30–32,54,58,12,65,67,29,76]
AI4	A priori algorithm[85]
AI5	K-Nearest neighbor classifier[2,20]
AI6	Radial basic function neighbor classifier[40,51]
AI7	Naive Bayes[2,4,61,20,26,37,39,58,65,67]
AI8	Classification association rules mining[43]
AI9	Fuzzy inference[28]
AI10	Rule induction[12]

these are frequently used techniques for classifications based on data characteristics, and are flexible in the use of categorical and continuous predictor variables[34].

On the other hand, regarding the accuracy of data mining techniques, the authors considered metrics such as sensibility, specificity, and accuracy. Of these, accuracy is determined by the ratio of True Positives (TP) to True Negatives (TN) among the total of registers, as formulated in equation (1).

$$\frac{TP+TN}{TP+TN+FP+FN}*100 \quad (1)$$

where, FP is the number of false positives and FN the number of false negatives. Tables 12 and 13 report the accuracy levels of the data mining techniques that reached a ratio higher than 60% and have a dataset composed of a number higher than 100 students.

The results show that the most accurate techniques are the Decision Tree Classifier, with the classifiers C4.5, ID3, and CART, reaching an accuracy of 98%, 97.5%, and 97%, respectively. The results evidence that the most accurate technique is Linear Regression (87.8%). However, these results cannot be generalized, as they depend on the dataset and the considered variables.

e) Q4: What tools are used?

We identified four tools in studies with artificial intelligence techniques, and seven tools in those using statistical methods; these are presented in Tables 14 and 15, respectively. The results highlight that the most widely used tools are WEKA and SPSS Modeler, most likely due to their wide variety of automatic learning algorithms for data mining tasks, flexibility in predictive modeling, and their facilities and functionalities[26].

Table 11. Statistical techniques

IID	Technique
ES1	Logistical regression[7,25,63,28,69,34,39,40,54,56,58,59,33,62,73–75,32,82,84]
ES2	Lineal regression[83,60,38,47–50,52,57,27,70–72,77–81]
ES3	Discriminant analysis[24]
ES4	Probit analysis[5]

4. Discussion

Of the 67 studies identified on university student dropout prediction, 18% contemplate the pre-processing phase. Therefore, this underlines the importance of this phase in obtaining variable properties, solving data anomalies, and increasing accuracy rates. We found that 90% of the studies regarding dropout prediction contemplate factor dimension, which evidences its relevance to the scientific community. Age, gender, ethnicity, and entrance exam performance are the most commonly used factors and correspond to the personal dimension. Although the total factors are wide-ranging, their behavior changes from one context to another; therefore, there is much controversy over which factors prove to be most efficient in university dropout prediction. With respect to factor selection techniques, 34% of studies used descriptive statistics and 7% used principal components analysis. This is one of the most relevant phases when predicting dropout due to its reduction in variable dimensionality. Thus, it allows us to adequately select the most predominant factors used as input variables in dropout prediction models. With regards to the techniques used to predict dropout, currently, statistical techniques are most commonly used. However, these are gradually being replaced by artificial intelligence techniques, since the latter present higher accuracy rates. Nevertheless, these rates vary according to the factors and educational context, the educational environment, and the theoretical framework of the analysis.

5. Conclusions

This study presents a systematic literature review on the aspects of data mining considered for predicting university dropout. We identified 1,681 primary studies related to the topic, from amongst which 67 documents were selected according to the established inclusion and exclusion criteria, identifying five important dimensions: factors, pre-processing techniques, factor selection techniques, prediction, and tools. This study makes an

Table 12. Accuracy of artificial intelligence techniques

Dataset size	Techniques	Accuracy (%)
200	Feed forward neural network[19]	82
	Probabilistic ensemble PESFAM[19]	62
	SEDM[19]	94
193	Feed forward neural network[41]	84
	Support vector machine[41]	83
	Probabilistic ensemble simplified fuzzy ARTMAP[41]	97
170	Naive Bayes[65]	81
	J48[65]	70
240	ID3[31]	92.50
	ID3 (Renyi)[31]	97.50
150	Support vector machine[12]	89.84
	Decision tree classifier[12]	86.32
	Rule induction[12]	81.98
3,200	Naive Bayes[67]	85
	Artificial neural networks[67]	62
	Decision trees and random forest[67]	63
62,375	Artificial neural network[20]	84
	Decision tree classifier[20]	82
	Bayesian networks[20]	76
300	C4.5[4]	98
	CART[4]	97
	Logistic regression[4]	86
775	Excalibur (J48)[61]	80
	SNA (PART)[61]	92
3,617	General Bayesian network[26]	89
	C4.5[26]	86
21,654	Artificial neural networks[63]	85
	Support vector machine[63]	90
	Decision tree classifier[63]	89
	Logistic regression[63]	80
	Logistic regression[39]	84
	Naive Bayes[39]	83
	Support vector machine[39]	67
	Decision tree classifier[39]	83
128	Decision tree classifier[58]	84
	Logistic regression[58]	84
	Naive Bayes[58]	82
	Artificial neural network[58]	82
189	K-Nearest neighbor[2]	87
	Decision tree classifier[2]	79
	Naive Bayes[2]	76
	Artificial neural network[2]	73
32,538	Logistic regression[34]	66
	Random forest[34]	62
	K-Nearest neighbor[34]	64
	ID3[30]	90.90
	C4.5[30]	89.09
	CART[30]	86.06
	ADT[30]	87.27
200	K-Nearest neighbor[51]	74
	Radial basis function[51]	70
	Support vector machine[51]	79
	Support vector machine[32]	65
	Logistic regression[32]	65
	Random forest[32]	86
	Gradient boosting decision tree[32]	88

inventory of 112 factors that influence dropout prediction. These factors were classified into five dimensions: personal, academic, economic, social, and institutional; the most commonly studied was the personal dimension, which considers factors such as age, ethnicity and gender. Furthermore, we identified ten pre-processing techniques, the most widely used being normalization and discretization. There were ten techniques for factor selection, of which descriptive statistics and Principal Component Analysis were the most referenced. Additionally, four-

Table 13. Accuracy of statistical techniques

Dataset Size	Technique	Accuracy (%)	Reference
237	Logistic regression	71.80	[73]
6,733		56.60	[59]
293		85.50	[69]
1,064		85.80	[56]
588	Linear regression	87.80	[50]
37,006		69.10	[78]
134		81.30	[27]
209	Discriminant analysis	78.20	[24]

Table 14. Tools used in studies applying artificial intelligence techniques

Tools	AI1	AI2	AI3	AI7	AI9
WEKA	46,3,61,36,37,58,29	63,61,36,58,29	4,63,61,37,30,58,65,67	4,61,37,65,67	28
SPSS Modeler	63,76	63,76	63		
Matlab	41,49	41	49		
SAS Enterprise	49		49		
Rapid Miner		12	12,67	67	

Table 15. Tools used in studies applying statistical techniques

Tools	S1	S2	S3
WEKA	28,63,58	27	
SPSS Modeler	7,63,33	83,38,27	
Matlab		49	85
R	84		
ISFE SYSTEM	77		
SAS Enterprise		49	
Excel	5		

teen techniques were identified for dropout prediction, and these were classified into statistics and artificial intelligence. The statistical techniques presented a higher frequency of use, while the artificial techniques presented greater accuracy rates. Finally, there are many data mining tools, of which the most used are WEKA and SPSS Modeler.

Consequently, it is clear that university dropout prediction is of interest to the scientific community, evidenced by the large volume of works on the topic, and its socio-economic impact. To address the problem of dropout, highly accurate techniques are being developed, however we cannot identify one technique that is clearly superior, for prediction accuracy depends mainly on the context, data and technique characteristics; any potential alternative must consider these factors.

6. References

1. Márquez-Vera C, Morales CR, Soto SV. Predicting school failure and dropout by using data mining techniques. IEEE Revista Iberoamericana de Tecnologias del Aprendizaje 2013; 8(1):7–14. https://doi.org/10.1109/RITA.2013.2244695
2. Yukselturk E, Ozekes S, Türel YK. Predicting dropout student: an application of data mining methods in an online education program. European Journal of Open, Distance and E-learning. 2014; 17(1):118–33. https://doi.org/10.2478/eurodl-2014-0008
3. Lin SH. Data mining for student retention management. Journal of Computing Sciences in Colleges. 2012; 27(4): 92–9.
4. Hu YH, Lo CL, Shih SP. Developing early warning systems to predict students' online learning performance. Computers in Human Behavior. 2014; 36:469–78. https://doi.org/10.1016/j.chb.2014.04.002
5. Jia P, Malone T. Using predictive modelling to identify students at risk of poor university outcomes. Higher Education. 2015; 70(1):127–49. https://doi.org/10.1007/s10734-014-9829-7
6. Lye CT, Ng LN, Hassan MD, Goh WW, Law CY, Ismail N. Predicting Pre-university student's Mathematics achievement. Procedia-Social and Behavioral Sciences. 2010; 8:299–306. https://doi.org/10.1016/j.sbspro.2010.12.041
7. Wray J, Barrett D, Aspland J, Gardiner E. Staying the course: Factors influencing pre-registration nursing student progression into Year 2-A retrospective cohort study. International Journal of Nursing Studies. 2012; 49(11):1432–42. https://doi.org/10.1016/j.ijnurstu.2012.06.006. PMid:22770946
8. Rodríguez-Gómez D, Feixas M, Gairín J, Mu-oz JL. Understanding Catalan University dropout from a comparative approach. Procedia-Social and Behavioral Sciences. 2012; 46:1424–9. https://doi.org/10.1016/j.sbspro.2012.05.314
9. Cantú-Martínez PC. Educación ambiental y la escuela como espacio educativo para la promoción de la sustentabilidad. Revista Electrónica Educare. 2014; 18(3):39–52. https://doi.org/10.15359/ree.18-3.3
10. Sittichai R. Why are there dropouts among university students? Experiences in a Thai University. International Journal of Educational Development 2012; 32(2):283–9. https://doi.org/10.1016/j.ijedudev.2011.04.010
11. Rebbapragada S, Basu A, Semple J. Data mining and revenue management methodologies in college admissions. Communications of the ACM. 2010; 53(4):128–33. https://doi.org/10.1145/1721654.1721690
12. Romero C, Ventura S. Educational data mining: A review of the state of the art. IEEE Transactions on Systems, Man, and

Cybernetics. 2010; 40(6):601–18. https://doi.org/10.1109/TSMCC.2010.2053532
13. Papamitsiou Z, Economides AA. Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. Journal of Educational Technology and Society. 2014; 17(4):49–64.
14. Bakhshinategh B, Zaiane OR, ElAtia S, Ipperciel D. Educational data mining applications and tasks: A survey of the last 10 years. Education and Information Technologies. 2018; 23(1):537–53. https://doi.org/10.1007/s10639-017-9616-z
15. Kitchenham B. Procedures for performing systematic reviews. Keele, UK, Keele University. 2004; 33:1–26.
16. Girón F. Factores de Riesgo que Ocasionaron la Deserción de Estudiantes de la Facultad de Ingeniería de la Universidad Rafael Ladívar. Tesis de maestría). Universidad Rafael Landívar; 2014. p. 1–71.
17. Romero C, Ventura S, Pechenizkiy M, Baker RS. Handbook of educational data mining. CRC Press; 2010. p. 1–526. https://doi.org/10.1201/b10274. PMCid:PMC3568769
18. Rivero Pérez JL. Técnicas de aprendizaje automático para la detección de intrusos en redes de computadoras. Revista Cubana de Ciencias Informáticas. 2014; 8(4):52–73.
19. Lara JA, Lizcano D, Martínez MA, Pazos J, Riera T. A system for knowledge discovery in e-learning environments within the European Higher Education Area-Application to student data from Open University of Madrid, UDIMA. Computers and Education. 2014; 72:23–36. https://doi.org/10.1016/j.compedu.2013.10.009
20. Tan M, Shao P. Prediction of student dropout in e-Learning program through the use of machine learning method. International Journal of Emerging Technologies. 2015; 10(1):11–17. https://doi.org/10.3991/ijet.v10i1.4189
21. Maimon O, Browarnik A. NHECD-Nano health and environmental commented database. Data mining and knowledge discovery handbook, Springer; 2009. p. 1221–41. https://doi.org/10.1007/978-0-387-09823-4_64
22. Yasmin D. Application of the classification tree model in predicting learner dropout behaviour in open and distance learning. Distance Education. 2013; 34(2):218–31. https://doi.org/10.1080/01587919.2013.793642
23. Nistor N, Neubauer K. From participation to dropout: Quantitative participation patterns in online university courses. Computers and Education. 2010; 55(2):663–72. https://doi.org/10.1016/j.compedu.2010.02.026
24. Sangodiah A, Beleya P, Muniandy M, Heng LE, Spr CR. Minimizing student attrition in higher learning institutions in Malaysia using support vector machine. Journal of Theoretical and Applied Information Technology. 2015; 71(3):1–9.
25. Arulampalam W, Naylor RA, Smith JP. Dropping out of medical school in the UK: Explaining the changes over ten years. Medical Education. 2007; 41(4):385–94. https://doi.org/10.1111/j.1365-2929.2007.02710.x PMid:17430284
26. Xing W, Chen X, Stein J, Marcinkowski M. Temporal predication of dropouts in MOOCs: Reaching the low hanging fruit through stacking generalization. Computers in Human Behavior. 2016; 58:119–29. https://doi.org/10.1016/j.chb.2015.12.007
27. Chen R. Institutional characteristics and college student dropout risks: A multilevel event history analysis. Research in Higher Education. 2012; 53(5):487–505. https://doi.org/10.1007/s11162-011-9241-4
28. Bayer J, Bydzovská H, Géryk J, Obsivac T, Popelinsky L. Predicting drop-out from social behaviour of students. International Educational Data Mining Society; 2012. p. 1–7.
29. Hovdhaugen E. Transfer and dropout: Different forms of student departure in Norway. Studies in Higher Education. 2009; 34(1):1–17. https://doi.org/10.1080/03075070802457009
30. Heredia D, Amaya Y, Barrientos E. Student dropout predictive model using data mining techniques. IEEE Latin America Transactions. 2015; 13(9):3127–34. https://doi.org/10.1109/TLA.2015.7350068
31. Pal S. Mining educational data to reduce dropout rates of engineering students. International Journal of Information Engineering and Electronic Business. 2012; 4(2):1–7. https://doi.org/10.5815/ijieeb.2012.02.01
32. Sivakumar S, Venkataraman S, Selvaraj R. Predictive modeling of student dropout indicators in educational data mining using improved decision tree. Indian Journal of Science and Technology. 2016; 9(4):1–5. https://doi.org/10.17485/ijst/2016/v9i4/87032
33. Liang J, Yang J, Wu Y, Li C, Zheng L. In Big data application in education: dropout prediction in edx MOOCs. IEEE Second International Conference on Multimedia Big Data (BigMM); 2016. p. 440–43. https://doi.org/10.1109/BigMM.2016.70 PMCid:PMC5052158
34. Herzog S. Measuring determinants of student return vs. dropout/stopout vs. transfer: A first-to-second year analysis of new freshmen. Research in Higher Education. 2005; 46(8):883–928. https://doi.org/10.1007/s11162-005-6933-7
35. Predicting student dropout in higher education [Internet]. [cited 2016 Jun 20]. Available from: https://arxiv.org/abs/1606.06364.
36. Wood L, Kiperman S, Esch RC, Leroux AJ, Truscott SD. Predicting dropout using student-and school-level factors: An ecological perspective. School Psychology Quarterly. 2017; 32(1):1–35. https://doi.org/10.1037/spq0000152. PMid:27030991
37. Natek S, Zwilling M. Student data mining solution–knowledge management system related to higher education

institutions. Expert Systems with Applications. 2014; 41(14):6400–7. https://doi.org/10.1016/j.eswa.2014.04.024
38. Lam-On N, Boongoen T. Using cluster ensemble to improve classification of student dropout in Thai university. International Conference on Soft Computing and Intelligent Systems (SCIS) and 15th International Symposium on Advanced Intelligent Systems (ISIS); 2014. p. 452–7. https://doi.org/10.1109/SCIS-ISIS.2014.7044875
39. Paura L, Arhipova I. Cause analysis of students' dropout rate in higher education study program. Procedia-Social and Behavioral Sciences. 2014; 109:1282–6. https://doi.org/10.1016/j.sbspro.2013.12.625
40. Li W, Gao M, Li H, Xiong Q, Wen J, Wu Z. Dropout prediction in MOOCs using behavior features and multi-view semi-supervised learning. International Joint Conference on Neural Networks (IJCNN); 2016. p. 3130–7. https://doi.org/10.1109/IJCNN.2016.7727598
41. Hoffait AS, Schyns M. Early detection of university students with potential difficulties. Decision Support Systems. 2017; 101:1–11. https://doi.org/10.1016/j.dss.2017.05.003
42. Lykourentzou I, Giannoukos I, Nikolopoulos V, Mpardis G, Loumos V. Dropout prediction in e-learning courses through the combination of machine learning techniques. Computers and Education. 2009; 53(3):950–65. https://doi.org/10.1016/j.compedu.2009.05.010
43. Janosz M, Archambault I, Morizot J, Pagani LS. School engagement trajectories and their differential predictive relations to dropout. Journal of social Issues. 2008; 64(1): 21–40. https://doi.org/10.1111/j.1540-4560.2008.00546.x
44. Badr G, Algobail A, Almutairi H, Almutery M. Predicting students' performance in university courses: A case study and tool in KSU mathematics department. Procedia Computer Science. 2016; 82:80–9. https://doi.org/10.1016/j.procs.2016.04.012
45. Levy Y. Comparing dropouts and persistence in e-learning courses. Computers and Education. 2007; 48(2):185–204. https://doi.org/10.1016/j.compedu.2004.12.004
46. Oeda S, Hashimoto G. Log-data clustering analysis for dropout prediction in beginner programming classes. Procedia Computer Science. 2017; 112:614–21. https://doi.org/10.1016/j.procs.2017.08.088
47. Huang S, Fang N. Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. Computers and Education. 2013; 61:133–45. https://doi.org/10.1016/j.compedu.2012.08.015
48. Koonce DA, Hening DA. Data imputation to identify potential dropouts. Proceedings, Institute of Industrial and Systems Engineers (IISE); 2009. p. 1–246.
49. Jadrić M, Garača Ž, Čukušić M. Student dropout analysis with application of data mining methods. Management: Journal of Contemporary Management Issues. 2010; 15(1):31–46.
50. Park JH, Choi HJ. Factors influencing adult learners' decision to drop out or persist in online learning. Journal of Educational Technology and Society. 2009; 12(4):207–17.
51. Dewan MAA, Lin F, Wen D. Predicting dropout-prone students in e-learning education system. IEEE 12th International Conference on Ubiquitous Intelligence and Computing and 2015 IEEE 12th International Conference on Autonomic and Trusted Computing and 2015 IEEE 15th International Conference on Scalable Computing and Communications and its Associated Workshops (UIC-ATC-ScalCom); 2015. p. 1735–40. https://doi.org/10.1109/UIC-ATC-ScalCom-CBDCom-IoP.2015.315
52. Duque LC. A framework for analysing higher education performance: students' satisfaction, perceived learning outcomes, and dropout intentions. Total Quality Management and Business Excellence. 2014; 25(1–2):1–21. https://doi.org/10.1080/14783363.2013.807677
53. Onah DF, Sinclair J, Boyatt R. Dropout rates of massive open online courses: Behavioural patterns. EDULEARN14 Proceedings; 2014. p. 5825–34.
54. Fei M, Yeung DY. Temporal models for predicting student dropout in massive open online courses. IEEE International Conference on Data Mining Workshop (ICDMW); 2015. p. 256–63. https://doi.org/10.1109/ICDMW.2015.174
55. Saranya A, Rajeswari J. Enhanced prediction of student dropouts using fuzzy inference system and logistic regression. ICTACT Journal on Soft Computing. 2016; 6(2):1–6.
56. Reschly AL, Christenson SL. Prediction of dropout among students with mild disabilities: A case for the inclusion of student engagement variables. Remedial and Special Education. 2006; 27(5):276–92. https://doi.org/10.1177/07419325060270050301
57. Alkan N. Humor, loneliness and acceptance: Predictors of university drop-out intentions. Procedia-Social and Behavioral Sciences. 2014; 152:1079–86. https://doi.org/10.1016/j.sbspro.2014.09.278
58. Sultana S, Khan S, Abbas MA. Predicting performance of electrical engineering students using cognitive and non-cognitive features for identification of potential dropouts. International Journal of Electrical Engineering Education. 2017; 54(2):105–118. https://doi.org/10.1177/0020720916688484
59. Chen R, DesJardins SL. Exploring the effects of financial aid on the gap in student dropout risks by income level. Research in Higher Education. 2008; 49(1):1–18. https://doi.org/10.1007/s11162-007-9060-9
60. Yi H, Zhang L, Yao Y, Wang A, Ma Y, Shi Y, Chu J, Loyalka P, Rozelle S. Exploring the dropout rates and causes of dropout in upper-secondary Technical and Vocational Education and Training (TVET) schools in China. International Journal of Educational Development. 2015; 42:115–23. https://doi.org/10.1016/j.ijedudev.2015.04.009

61. do Nascimento RLS, das Neves Junior RB, de Almeida Neto MA, de Araújo Fagundes RA. Educational data mining: An application of regressors in predicting school dropout. International Conference on Machine Learning and Data Mining in Pattern Recognition; 2018. p. 246–57. https://doi.org/10.1007/978-3-319-96133-0_19
62. Tumen S, Shulruf B, Hatti J. Student pathways at the university: Patterns and Predictors of Completion. Studies in Higher Education. 2008; 33(3):233–52. https://doi.org/10.1080/03075070802049145
63. Thammasiri D, Delen D, Meesad P, Kasap N. A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. Expert Systems with Applications. 2014; 41(2):321–30. https://doi.org/10.1016/j.eswa.2013.07.046
64. Vogel C, Hochberg J, Hackstein S, Bockshecker A, Bastiaens TJ, Baumöl U. Dropout in distance education and how to prevent it, EdMedia+ innovate learning. Association for the Advancement of Computing in Education (AACE); 2018. p. 1788–99.
65. Al-barrak MA, Al-razgan MS. Predicting students' performance through classification: A case study. Journal of Theoretical and Applied Information Technology. 2015; 75(2):167–75.
66. Iepsen EF, Bercht M, Reategui E. In Detection and assistance to students who show frustration in learning of algorithms. IEEE Frontiers in Education Conference (FIE); 2013. p. 1183–9. https://doi.org/10.1109/FIE.2013.6685017
67. Guarín CEL, Guzmán EL, González FA. A model to predict low academic performance at a specific enrollment using data mining. IEEE Revista Iberoamericana de tecnologias del Aprendizaje. 2015; 10(3):119–25. https://doi.org/10.1109/RITA.2015.2452632
68. Martinho VR, Nunes C, Minussi CR. In Prediction of school dropout risk group using neural network. 2013 Federated Conference on Computer Science and Information Systems; 2013. p. 111–14.
69. Duarte R, Ramos-Pires A, Gonçalves H. Identifying at-risk students in higher education. Quality Control and Applied Statistics. 2015; 60(5):557–8.
70. Willging PA, Johnson SD. Factors that influence students' decision to drop out of online courses. Journal of Asynchronous Learning Networks. 2009; 13(3):115–27.
71. González-Flores M, Heracleous M, Winters P. Leaving the safety net: an analysis of dropouts in an urban conditional cash transfer program. World Development. 2012; 40(12): 2505–21. https://doi.org/10.1016/j.worlddev.2012.05.020
72. Belo P, Oliveira C. The relation between experiences and expectations with university dropout. Procedia-Social and Behavioral Sciences. 2015; 187:98–101. https://doi.org/10.1016/j.sbspro.2015.03.019
73. Stratton LS, O'Toole DM, Wetzel JN. A multinomial logit model of college stopout and dropout behavior. Economics of Education Review. 2008; 27(3):319–31. https://doi.org/10.1016/j.econedurev.2007.04.003
74. Arulampalam W, Naylor RA, Smith JP. Effects of in-class variation and student rank on the probability of withdrawal: Cross-section and time-series analysis for UK university students. Economics of Education Review. 2005; 24(3): 251–62. https://doi.org/10.1016/j.econedurev.2004.05.007
75. Di Pietro G, Cutillo A. Degree flexibility and university drop-out: The Italian experience. Economics of Education Review. 2008; 27(5):546–55. https://doi.org/10.1016/j.econedurev.2007.06.002
76. Hershkovitz A, Nachmias R. Online persistence in higher education web-supported courses. The Internet and Higher Education. 2011; 14(2):98–106. https://doi.org/10.1016/j.iheduc.2010.08.001
77. Ćukušić M, Garača Ž, Jadrić M. Online self-assessment and students' success in higher education institutions. Computers and Education. 2014; 72:100–9. https://doi.org/10.1016/j.compedu.2013.10.018
78. Oseguera L, Rhee BS. The influence of institutional retention climates on student persistence to degree completion: A multilevel approach. Research in Higher Education. 2009; 50(6):546–69. https://doi.org/10.1007/s11162-009-9134-y
79. Willcoxson L, Cotter J, Joy S. Beyond the first-year experience: The impact on attrition of student experiences throughout undergraduate degree studies in six diverse universities. Studies in Higher Education. 2011; 36(3): 331–52. https://doi.org/10.1080/03075070903581533
80. Human-Vogel S, Rabe P. Measuring self-differentiation and academic commitment in University students: A case study of education and engineering students. South African Journal of Psychology. 2015; 45(1):60–70. https://doi.org/10.1177/0081246314548808
81. Melguizo T, Sanchez F, Velasco T. Credit for low-income students and access to and academic performance in higher education in Colombia: A regression discontinuity approach. World Development. 2016; 80:61–77. https://doi.org/10.1016/j.worlddev.2015.11.018
82. Arbiv DC, Meiran N. Performance on the antisaccade task predicts dropout from cognitive training. Intelligence. 2015; 49;25–31. https://doi.org/10.1016/j.intell.2014.11.009
83. Elffers L. Staying on track: behavioral engagement of at-risk and non-at-risk students in post-secondary vocational education. European Journal of Psychology of Education. 2013; 28(2):545–62. https://doi.org/10.1007/s10212-012-0128-3
84. Arifin MH. Exploring factors in contributing student progress in the Open University. International Journal of Information and Education Technology. 2016; 6(1):1–29. https://doi.org/10.7763/IJIET.2016.V6.653
85. Aziz AA, Idris WMRW, Hassan H, Jusoh JA, Emran NA. Implementing Aproiri Algorithm for Predicting Result Analysis. GSTF Journal on Computing (JOC). 2018; 2(4):87–92.