Revista ELectrónica de Investigación y EValuación Educativa

RELIEVE 25 YEARS AÑOS

ISSN: 1134-4032

e-Journal of Educational Research, Assessment and Evaluation

University dropout: Prevention patterns through the application of educational data mining

Deserción escolar universitaria: Patrones para prevenirla aplicando minería de datos educativa

Urbina-Nájera, A.B. [(1)]; **Camino-Hampshire, J.C.**[(2)], **& Cruz Barbosa, R.**[(3)]

[(1)] UPAEP-University (Mexico); [(2)] Accenture (México) & UPAEP-University (Mexico); [(3)] Mixteca´s Technology University (Mexico)

Abstract

Recently, the use of educational data mining techniques has gained great relevance when applied to performance prediction, creation of predictive retention models, behaviour profiles and school failure, amongst others. For the present paper we applied an attribute selection algorithm to identify the most important factors influencing drop out decision. Decision trees were used to define patterns that can alert an imminent dropout. A tool was adapted and administered online to 300 students from public HEIs, and 200 students from private HEIs currently enrolled on a higher education program. By means of the attribute selection algorithm, 27 relevant factors were found. Within the three main factors, the lack of counselling, an adequate student environment and academic follow-up were recognized, whilst, 7 patterns were found through the decision tree. These included factors such as: student environment, insufficient financial support, experience of an uncomfortable situation and place of career choice, amongst others. Finally, it has been seen that school drop-out does not depend on a single factor but is multifactorial. It is imperative to expand the sample to include other cities. This will enable various algorithms to be applied, providing greater information and leading to the establishment of accurate mechanisms for reducing university drop-out rates, according to the characteristics of the student population in each region.

Keywords: Student environment; Computer learning; Decision trees, Counseling; Feature selection

Resumen

Recientemente, el uso de técnicas de minería de datos educativa ha cobrado gran relevancia al aplicarlas en la predicción del desempeño, creación de modelos predictivos de retención, perfiles de comportamiento, fracaso escolar, entre otros. En este trabajo se presenta la aplicación del algoritmo selección de atributos para identificar los factores más importantes que inciden en la decisión de desertar; también, se utilizan árboles de decisión para definir patrones que pueden alertar una inminente deserción. Se adaptó un instrumento y se administró vía web a 300 estudiantes de IES pública y 200 estudiantes de IES privada actualmente inscritos en algún programa de nivel superior. Mediante el algoritmo selección de atributos se encontraron 27 factores relevantes, dentro de los tres factores principales se reconocen la falta de asesorías, la falta de un ambiente estudiantil adecuado y la falta de seguimiento académico, mientras que, por medio del árbol de decisión se encontraron 7 patrones, en donde uno de ellos incluye factores como: ambiente estudiantil, apoyos financieros insuficientes, experiencia de una situación incómoda, lugar que ocupa la elección de la carrera, entre otros. Finalmente, se ha visto que la deserción escolar no depende de un solo factor, sino que es multifactorial y que es imperativo ampliar la muestra a otras ciudades de manera que se puedan aplicar diversos algoritmos que proporcionen mayor información que conduzcan al establecimiento de mecanismos certeros para disminuir los índices de deserción universitaria en función de las características de la población estudiantil según la región.

Palabras clave: Ambiente estudiantil, Aprendizaje computacional, Árboles de decisión, Asesoramiento, Selección de atributos

Received/*Recibido* 2019 November 04	**Approved /*Aprobado*** 2020 April 20	**Published/*Publicado*** 2020 June 30

Corresponding author / Autor de contacto: **Urbina-Nájera, A.B.** UPAEP-Universidad. 17 sur #901.
Barrio de Santiago. 72410. Puebla, Pue. (México) abunajera@gmail.com

Higher education is the engine that drives the growth of modern economies as it equips students with specific competencies. These include technical, professional, human and disciplinary skills which qualify students for diverse job functions (OECD, 2019). It is for this reason that education and the development of competencies are the pillars on which nations should build their growth, productivity and future prosperity (OECD, 2017). In this sense, the Mexican higher education system offers a wide range of programs which have experienced rapid development over recent decades. In particular, 89% of students enrolled on undergraduate or engineering studies in 2015, representing a higher percentage than the OECD (Organization for Economic Co-operation and Development) average of 61%. However, in comparison to the OECD average of 37% (OECD, 2019), Mexico has a lower proportion of adults aged 25-64 with a higher education degree (17%).

Despite these figures, Mexico has made significant progress in increasing educational attainment in higher education, with the proportion of young adults having completed higher education rising from 17% to 23% over the past 16 years (OECD, 2019). This slow growth is due to school drop-out. This aspect of education has received special attention over the past decade in all nations. According to data reported by the World Bank (2018), half of all students between the ages of 25 and 29 complete their university studies, with 50% of dropouts occurring in the first year of university courses. According to data from the National Institute of Statistics and Geography (INEGI, 2018), only 8 out of every 100 Mexican university students finish their studies. Finally, the Ministry of Public Education (SEP) reported that in the 2017-2018 school year there was an 8.4% increase in dropout. This compared to a 7.2% increase seen in the 2016-2017 school year (Secretaría de Educación Pública, 2019).

Dropout is understood to be an act that leads to withdrawing from studies or incorporation into another institution. School dropout is characterised by different priorities superseding that of studying, attending to obligations that are met by various human needs. In other words, dropout is determined from statistical figures related to the number of students who leave school from one school year to another (Ministry of Public Education, 2019). In this way, a dropout is understood as an individual who has abandoned his or her studies and obligations as a student, losing his or her status as a student and the rights acquired through his or her enrolment at the educational centre (Universidad Tecnológica de Tabasco, 2019).

Various studies show that the causes that lead to dropping out of school are multifactorial and range from personal, family and economic reasons to political, cultural and institutional aspects. In the present work, the term dropout is used to identify those individuals who drop out or decide not to resume their studies at any period of their academic training.

Recently, the use of educational data mining techniques has become relevant in the analysis of various educational aspects such as school dropout. Its application aims to trace the behaviour of students and discover, in a timely fashion, behaviour change linked to academic aspects which can predict, for example, an imminent dropout or desertion. This technique has also been used to predict student behaviour in order to make recommendations on the learning-teaching process, performance, management of activities, etc. Similarly, it has been used to find hidden patterns in at-risk students at risk, boost retention and avoid debt. Educational data mining seeks to create methods to explore the unique types of data that come from learning environments in order to solve and improve educational processes in an automated way (Romero & Ventura 2007).

In this way, the objective of the present study is to make use of this computational technique by applying the attribute selection algorithm and the decision tree algorithm. The purpose of this is to identify the main factors that influence university desertion, whilst at the same time finding patterns to prevent it. It uses a dataset of 500 records recovered through a survey administered to students enrolled in public and private Higher Education Institutions (HEI) in the City of Puebla.

For this purpose, a brief review of the literature on university desertion is presented based on two aspects: traditional methodologies (quantitative and qualitative) and the application of educational data mining. Next, educational data mining is defined, highlighting its capacity to identify new and non-trivial patterns to solve and improve educational processes. At the same time, the methodology implemented is shown, which is based on the application of the knowledge discovery process in databases. Finally, we present the results that have made it possible to identify the patterns that influence school drop-out rates, and conclude by suggesting actions which will favour timely prevention and help to reduce local and national drop-out rates.

Educational data mining

Educational Data Mining ([EDM]) is an emerging discipline. It is focused on creating methods to explore the unique types of data that come from educational environments in order to solve and improve educational processes in an automated way. EDM methods are drawn from a variety of areas including data mining, computer learning, psychometrics, statistics, information visualization and computer modelling (Romero and Ventura 2007). In the present day, there are many methods (algorithms) which have been accurately applied to various real-world problems. These algorithms include decision trees, vector support machines, artificial neural networks, Bayesian learning, instance-based methods, kernel methods, etc. In this section, a brief description of decision trees is given, as well as the metrics used to evaluate their performance.

Decision Trees (DT)

Decision trees are located within a branch of computer learning called symbolic learning, which also includes decision rule models, which are closely related to trees. Decision tree learning is a technique for analysing sequential decisions based on the use of outcomes and associated probabilities. Mitchel (1997) defines it as "a method of approximating an objective function of discrete values in which the objective function is represented by a decision tree”.

Learned trees can also be represented as a set of “If-then rules...", with these being one of the most widely used inductive learning methods in inference algorithms. Such representation is denoted by a decision node, a probability node and a branch. The decision node, represented by a square, indicates that a decision needs to be made at that point in the process. A probability node, represented by a rounded rectangle, indicates that a random event occurs at that point in the process. Finally, the branch shows the different paths that can be taken when a decision is made or a random event, represented by a line, occurs (Frank, Hall, Mark & Witten, 2016).

Most of the algorithms that have been developed for learning decision trees are variations of an algorithm that employs a top-down kernel. Specifically, this approach includes the ID3 algorithm and its successor C4.5, which were both developed by Quinlan in 1986 and 1993, respectively.

The ID3 (Induction Decision Trees) algorithm is a supervised learning system that builds decision trees from a set of examples. These examples, or tuples, are constituted by a set of attributes and a classifier or class. The domains of the attributes and classes must be discrete. In addition, the classes must be independent. In general, algorithms generate descriptions to classify each of the examples of the training set. The algorithm C4.5 or J48 (which is an extension of ID3) allows continuous values for the attributes to be worked with, separating the possible results into two branches. The trees it generates are less leafy because each leaf does not cover a particular class but a distribution of classes. This algorithm is popularly used in data mining because of its simplicity of interpretation and the visual representation is gives of results (Mitchell, 2000).

Other popular algorithms applied in this area are NBTree and Random Forest. The NBTree algorithm (Naive Bayes Tree) is considered a hybrid algorithm because of the tree that it generates. The leaves of this tree contain a Naive Bayes classifier built from the examples entered into the nodes. It is an efficient and effective learning algorithm, whilst also presenting predictive data as efficiently as the C4.5 algorithm. However, it is limited by the fact that it only represents a certain degree of separation between binary functions (Chen et al., 2017). Random Forest, on the other hand, is made up of a large number of individual decision trees which operate as a whole. In this case, each individual tree in the random forest shows a class prediction and the class with the most votes becomes the model prediction (Ustebay, Turgut and Ali, 2018).

Similarly, there are other algorithms such as CHAID or CART (none available in Weka). The former is commonly used to measure the degree of correlation between independent variables and class (Cha, Kim, Moon & Hong, 2017), whilst the latter is used for linear or multiple regression analysis (Sharma & Kumar, 2016).

On the other hand, in order to build a TD it is necessary to determine which attributes are the best, particularly, which attribute should be placed in the root node. Thus, entropy and information gain are used to answer these questions. According to Mitchel (1997), entropy is a measure that allows us to calculate the degree of uncertainty of a sample. If the sample is completely homogeneous, its entropy = 0, as opposed to an equally distributed sample, whose entropy = 1. In this sense, information gain is understood as the quality of a variable, i.e., information gain verifies how homogeneous the distribution of the class is before instantiating any variable. This approach is specifically used in the creation of a decision tree (Gupta, Rawat, Jain, Arora & Dhami, 2017).

Performance of decision trees

An algorithm must be analysed in order to determine resources use and, mainly, task performance for classifying, recognising, identifying, grouping and categorising, etc. There are measures to estimate the performance of an algorithm and this performance will depend on what measure is being used as a priority.

The most obvious criterion for estimating the performance of a classifier is its predictive accuracy in instances that are not observed. The number of unobserved cases is sometimes potentially large (if not infinite), so an estimate must be calculated in a test suite. This is commonly referred to as cross validation. Cross-validation is a technique used to evaluate the results of a statistical analysis and ensure that they are independent of the partitioning of training and test data. This method is very accurate as it is evaluated from k combinations of training and test data. In practice, the choice of the number of iterations depends on the size of the dataset. The most common selection process is the 10-fold cross-validation. In this case, if the sample is very large (k>10) then k = 3, and if the sample is very small, then the maximum value of k is taken (M.P. van der Aalst, 2011).

On the other hand, a confounding matrix, also called a prediction or classification matrix, is a visualisation tool used to obtain information about the actual and predicted classifications made by a classification system (Bird, Klein & Loper, 2009). Thus, the confounding matrix is a table where each cell [i,j] indicates how some instance was classified with respect to its pre-established class (Table 1). Accurately classified cases are found in the diagonal entries (in this case cells [a,d] of Table 1) because the predicted and real groups are the same. Elements found outside of the diagonal are poorly classified (Witten & Frank, 2005; Montero Lorenzo, 2007; Bird, Klein & Loper, 2009; Hamilton, 2009).

Table 1. Confusion matrix when there are two possible classification results: Negative and positive

		Prediction	
		Negative	**Negative**
Actual	**Negative**	a	b
	Positive	c	d

Frequently used metrics obtained from the confusion matrix are accuracy, precision, recall and F measurement, the concepts of which are described here: Accuracy is defined as the proportion of the total number of predictions that are correct; precision (consistency or confidence) is known as the proportion of the prediction of correct positive cases; recall (true positive, completeness or sensitivity) is interpreted as the number of cases that should have been recovered based on determined search criteria; finally, F-measure is a measure that combines accuracy with recall to produce a single score (Freitas, 2002; Witten and Frank, 2005; Bird, Klein & Loper, 2009; M. P. van der Aalst, 2011).

Feature selection

It is common to have a large number of attributes for each instance in a data set, however, not all of these may be relevant for characterising the object. In fact, using all of the attributes can, in many cases, cause a problem (Witten, Frank & Hall, 2011). In other words, a large number of attributes represents a large dimensional space, making it necessary to carry out a reduction of the dimensionality, selecting only a few attributes. This small set of attributes should retain as much information as possible for describing the examples (Bishop, 2007).

On the other hand, attribute selection consists of an evaluator and a search method. An evaluator (individual or subset) defines how the algorithms evaluate attributes and are classified into filter, wrapper and ranker. The first two generate a subset of attributes, whilst the third aspect generates a ranking of all the attributes. A search method must be selected in order to be able to execute an evaluator. In the case of filter and wrapper evaluators, the method consists of searching a space between the subsets of data using one of a number of methods. One example is the CfsSubsetEval method which evaluates a subset of attributes and considers the individual predictive ability of each variable, in addition to the degree of inter-variable redundancy. The ConsistencySubsetEval method determines a subset of attributes according to the level of consistency in the class values when projecting the training instances over the subset of attributes. The ClassifierSubsetEval method uses a classifier to estimate the subsets of attributes in the training data or in a separate test set. The WrapperSubsetEval method calculates the subsets of attributes using a classifier. It uses cross-validation to estimate the accuracy of the learning scheme in each set (Hall, 2011; Frank, Hall, Mark & Witten, 2016).

Similarly, ranker evaluator (considered as an individual evaluator) search methods evaluate individual attributes and include the following methods. The ChiSquaredAttributeEval method calculates the Chi-squared statistical value of each attribute with respect to its class and in this way obtains the level of correlation between the class and each attribute. The GainRatioAttributeEval method evaluates each attribute by measuring its profit ratio with respect to the class. The InfoGainAttributeEval method estimates attributes by measuring the gain in information afforded by each one with respect to its class. The OneRAttributeEval method measures the quality of each attribute using the OneR classifier. This uses least error estimates of attributes to predict and discretise numerical attributes (Frank, Hall, Mark & Witten, 2016).

The knowledge discovery process

Knowledge Discovery in Databases (KDD) focuses mainly on the following tasks: data collection, data pre-processing, attribute selection and application of computer learning algorithms. Each of these tasks is described below.

Data collection. Data collection methods include acquiring and storing new observations, consulting existing databases according to the problem, and, if necessary, performing data combinations (Han, Kamber & Pei, 2011). In the particular case of the present work, data were collected in a digital format, the process for which is described in the following sections.

Pre-processing of the data. Pre-processing consists of manipulating, enriching, reducing or transforming original data in order to make it more easily accessible later on (Han, Kamber &

Pei, 2011). Following this, the transformation phase involves combining data which reside in different sources in order to provide a unified view of these data. This enables data to be converted from a source format to a target format according to the tool with which it will be used, and enables it to be loaded without any reading problems (Witten & Frank, 2005).

Feature selection. The appropriateness of attribute selection depends on the improvement it affords in prediction, reductions in the training time needed to use the algorithm and reduction in storage space.

Application of computational learning algorithms. Computational learning (CA) is the branch of Artificial Intelligence that is dedicated to the study of agents/programs that learn or evolve based on their experiences in order to perform a given task better (Mitchell, 2000). Computational learning algorithm application represents the last phase of the knowledge discovery process. Its main objective is to use known evidence to be able to create a hypothesis and to be able to give an answer to new unknown situations. Thus, from this basis, algorithms that give answers to diverse situations must be selected.

As mentioned above, there are several types of decision tree algorithms commonly used in the area of data mining such as ID3, C4.5, NBTree, RandomForest, CHAID and CART. The present work used the C4.5 algorithm for the following reasons. It facilitates interpretation of the adopted decision, provides a high degree of knowledge understanding, explains behaviour with respect to specific decision-making tasks, reduces the number of independent variables and enables problems to be visually displayed (Yukselturk, Ozekes & Kılıç, 2014; Jadhav & Channe, 2016). It is also commonly used to identify the most important variables in a dataset, and has a lower error rate and greater precision than other approaches (Sharma & Kumar, 2016; Gupta, Rawat, Jain, Arora & Dhami, 2017). Further, algorithm attribute selection based on the ranker evaluator, and the GainRatioAttributeEval and InfoGainAttributeEval search methods is used because these approaches make it possible to obtain an ordered list of the most relevant attributes within analysed datasets.

Related works

In this section, a brief review of the literature on university dropout is presented based on two aspects: 1) Studies that used conventional methodologies (quantitative, qualitative, statistical methods or theoretical reflections) and 2) Studies related to the application of computer-based learning in various educational contexts.

Studies using statistical methods

While there is no single cause for dropping out of school, there are reasons behind the decision to drop out. In order to better understand potential causes, this section presents 15 related works which were selected using the following inclusion criteria: Studies conducted between the year 2000 and the present date; using statistical methods (mixed approaches, qualitative approaches, quantitative approaches); and. focused on higher education. Related studies carried out with masters and doctoral students were excluded. Table 2 shows the list of factors identified by the employed statistical methods in alphabetical order.

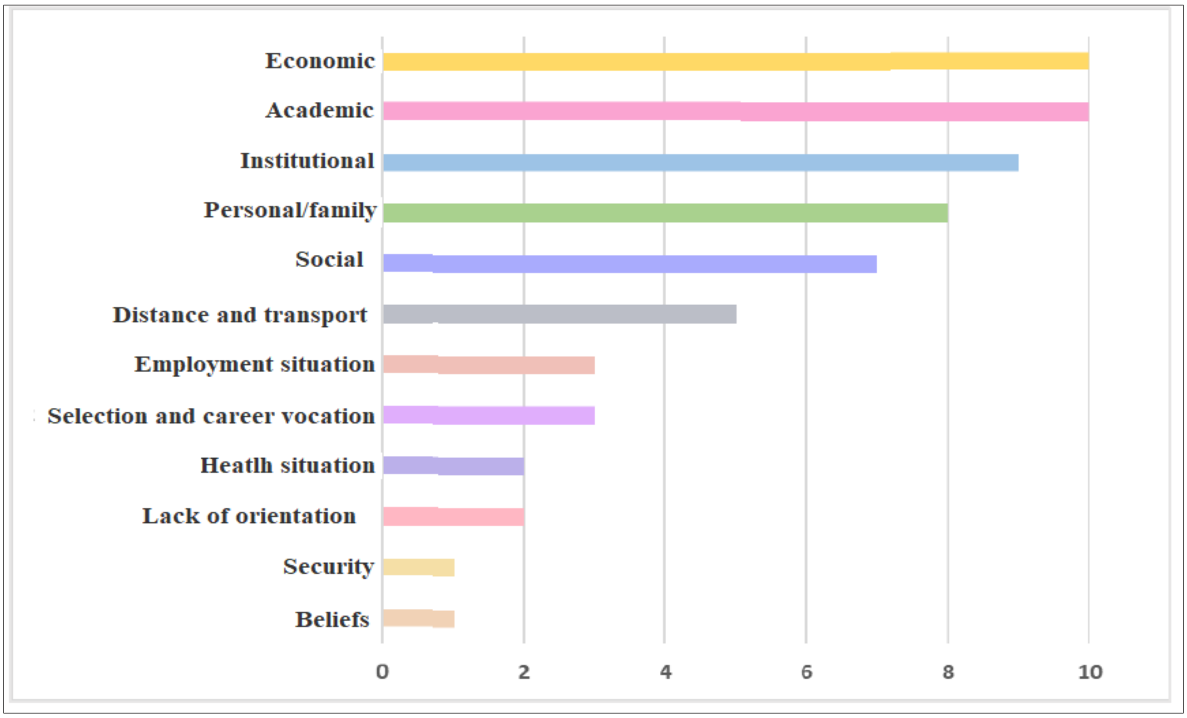
As detailed in Figure 1, the two most frequently appearing factors were economic and academic. The figures represent the number of authors who have mentioned each factor as the main explanatory cause in their studies (Table 2). Thus, to mention only a few, economic factors have been associated with lack of resources for tuition payments, lack of financial support, loss of employment of the parent or guardian and underestimation of costs. In the grouping of academic factors we find aspects related to insufficient or inadequate counselling, tutoring, accompaniment or guidance, poor personal academic performance, failure to comply with academic regulations, too high/low academic level of the university, and inappropriate course provision or schedule.

Table 2. University dropout factors identified with statistical methods from 2000-2016

Factors	Authors
Beliefs	Carvajal & Trejos (2016)
Distance and transport (coverage)	Vélez & López (2004); Lavado & Gallegos (2005); Sandoval (2001); Abarca & Sánchez (2005); Lugo (2013)
Academic factors	Ruíz (2009); Erazo, et. al. (2013); Londoño (2013); Rode, Bjornoy & Sogaard (2013); Lugo (2013); Fozdar, Kumar y Kannan (2006); Vries, León Arenas, Romero & Hernández, (2011); Carvajal & Trejos (2016)
Economic factors	Vélez & López (2004); Ruíz (2009); Lavado & Gallegos (2005); Erazo, et al. (2013); Sandoval (2001); Cabrera, Bethencourt, Pérez & González (2006); Londoño (2013); Rode, Bjornoy & Sogaard (2013); Lugo (2013) Carvajal y Trejos (2016)
Personal and family factors	Erazo, et al. (2013); Sandoval (2001); Cabrera, Bethencourt, Pérez & González (2006); Londoño (2013); Rode, Bjornoy & Sogaard (2013); Lugo (2013); Fozdar, Kumar & Kannan (2006); Carvajal & Trejos (2016)
Social factors	Vélez y López (2004); Ruíz (2009); Cabrera, Bethencourt, Pérez & González (2006); Londoño (2013); Rode, Bjornoy & Sogaard (2013); Lugo (2013); Carvajal & Trejos (2016)
Lack of orientation	Abarca & Sánchez (2005); Rode, Bjornoy & Sogaard (2013)
Institutional	De los Santos (2004); Ruíz (2009); Abarca & Sánchez (2005); Cabrera, Bethencourt, Pérez & González (2006); Londoño (2013); Rode, Bjornoy & Sogaard (2013); Fozdar, Kumar & Kannan (2006); Carvajal & Trejos (2016)
Security	Vélez & López (2004)
Selection/ career vocation	Abarca & Sánchez (2005); Lugo (2013); Vries, León, Romero & Hernández (2011)
Health situation	Erazo, et. Al. (2013); Lugo (2013)
Employment situation	Ruíz (2009); Rode, Bjornoy & Sogaard (2013); Vries, León, Romero & Hernández (2011)

Source: Own elaboration

Figure 1. Factors affecting university dropout identified by statistical methods used in a 2000-2016 review



Source: Own elaboration

Studies that applied computer-based learning

After the year 2000, other scientific techniques were used to identify causes leading to school dropout. This included not only university dropout, but also dropout that occurred in secondary/high school. In particular, the use of artificial intelligence techniques has become very important. Several studies describe the application of computer learning algorithms to resolve situations in the educational context. In this sense, the commonly used algorithms were identified.

- Näive Bayes (Kotsiantis, Pierrakeas & Pintelas [2003]; Dekker, Pechenizki & Vleeshouwers [2009]; Pal [2012]; Er [2012]; Yukselturk, Ozekes & Kılıç [2014]; Barbosa, Serra da Cruz & Zimbrão [2014]; Sara, Halland, Igel & Alstrup [2015]; Márquez-Vera, et al.) The highest reported prediction percentage was 83%, as reported by Kotsiantis, Pierrakeas & Pintelas (2003).
- Neural networks (Kotsiantis, Pierrakeas & Pintelas [2003], Delen [2011]; Yukselturk, Ozekes & Kılıç [2014]; Alkhasawneh & Hargraves [2014]; Barbosa, Serra da Cruz & Zimbrão [2014]). The highest reported prediction percentage was 81%, reported by Delen (2011) and Alkhasawneh & Hargraves (2014).
- Nearest Neighbours (K-NN) (Kotsiantis, Pierrakeas & Pintelas [2003]; Yukselturk, Ozekes & Kılıç [2014]; Marquez-Vera, et al. [2016]; Aulck, Velagapudi, Blumenstock & West [2017]); with the highest prediction percentage reported as 87% by Yukselturk, Ozekes & Kılıç (2014).
- Regression (linear and logistic) (Kotsiantis, Pierrakeas & Pintelas [2003], Delen [2011]; Aulck, Velagapudi, Blumenstock & West [2017]; Yamao, Saavedra, Campos Pérez & Huancas [2018]). The highest prediction percentage reported was 66.59%, reported by Aulck, Velagapudi, Blumenstock & West (2017).
- Support vector machines (Kotsiantis, Pierrakeas & Pintelas [2003]; Barbosa Manhães, Serra da Cruz & Zimbrão [2014]; Sara, Halland, Igel & Alstrup [2015]; Márquez-Vera, Cano, Romero & Mohammad Noaman [2016]; Yamao, Saavedra, Campos & Huancas [2018]). The highest prediction percentage reported was 87.39%, reported by Barbosa, Serra da Cruz & Zimbrão (2014).
- Feature selection (Márquez, Cano, Romero and Ventura [2012]; Alkhasawneh & Hargraves [2014]; Márquez-Vera, et al).

Of course, decision trees have also been used to resolve various situations in the educational context. Table 3 shows work from 2003 to 2019 related to the application of the decision tree algorithm in the solution of educational situations in higher education. The highest prediction percentage reported was 82.87%, reported by Yamao, Saavedra, Campos and Huancas (2018).

Table 3. Educational situations resolved by applying decision trees between 2003 and 2019.

Educational situations	Authors
Performance prediction	Kabra & Bichkar (2011); Vijayalakshmi & Kumar (2011); Márquez, Cano, Romero & Ventura (2012); Barbosa, Serra da Cruz & Zimbrão (2014); Al-Barrak & Al-Razgan (2016); Agaoglu (2016); Chiheb, Boumahdi, Bouarfa & Boukraa (2017); Yamao, Saavedra, Campos & Huancas (2018)
Enrolment in private HEIs	Estrada-Danell, Zamarripa-Franco, Zúñiga-Garay & Martínez-Trejo (2016)
Academic success	Morales & Parraga-Alava (2018)
Dropout prediction	Dekker, Pechenizki & Vleeshouwers (2009);Yukselturk, Ozekes & Kılıç (2014); Abu-Oda & El-Halees (2015); Márquez-Vera, et al. (2016); Sivakumar, Venkataraman & Selvaraj (2016);
Predictive models of college dropouts	Aulck, Velagapudi, Blumenstock & West (2017); Rodríguez-Maya, Lara-Álvarez, May-Tzuc & Suárez-Carranza (2017);
Retention	Raju y Schumacker (2015); Delen (2011); Kumar, Bharadwaj & Pal (2012);
Behaviour profiles	Guevara, et al. (2019)
School failure	Márquez, Romero & Ventura (2012)
At-risk students	Er (2012)
Decrease in dropout rate	Pal (2012)

Source: Own elaboration

Following the review presented in Table 3, it was observed that only three papers applied the attribute selection algorithm to identify factors influencing high school dropout. In the first of these papers, a dataset with 77 attributes (characteristics) related to 670 young people aged between 15-18 years was used. The authors applied 10 methods to rank predominant factors, considering only those that had a frequency greater than or equal to 2 and selecting only 15 attributes from the 77 total attributes. Of these, the following emerge according to their importance: evaluations in various knowledge areas (eight attributes), level of motivation, secondary education grades, age, number of siblings, group, smoking habits and average score on entrance exams (Márquez, Cano, Romero and Ventura, 2013)

In the second article, a dataset was used which contained 20 attributes taken from 1,966 young people. Attribute selection was applied to obtain the most important dropout factors: Gender, total credits pertaining to each course unit, credits achieved in each unit, general average and average in mathematics (Alkhasawneh and Hargraves, 2014).

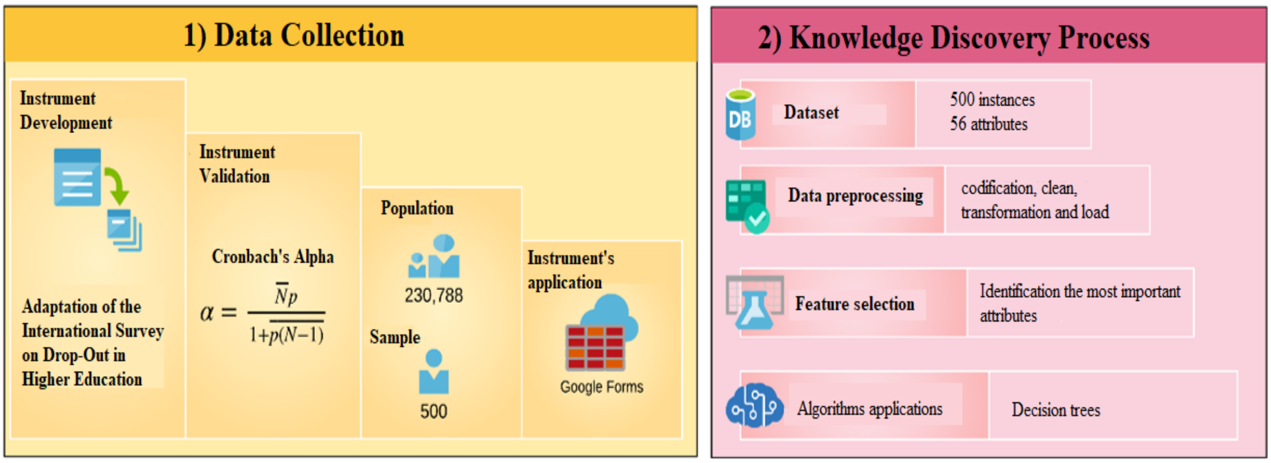
Finally, in the third article a dataset was used which consisted of 60 attributes retrieved from 419 high school students. Through application of the attribute selection algorithm the authors obtained the following main attributes: average grade in high school, group, number of students in the group, age, attendance, mother's educational level, distance, regular alcohol consumption, smoking habits, administrative sanctions, place used to study, level of motivation, and qualification in mathematics, social sciences and humanities (Márquez-Vera, et al., 2016)

Finally, from these findings it can be determined that the decision tree is the most commonly used algorithm for addressing situations in the educational context. Its highest associated prediction percentage was reported as 82.87%. Further, it is seen that the best performing algorithm for predicting educational events relate to vector support machines, with a prediction percentage of 87.39%. On the other hand, the algorithm with the lowest performance when predicting an educational situation is linear regression, with a percentage of 66.59%. These successes allow the present paper to identify the most important factors involved in university dropout from a dataset of 56 attributes using attribute selection, whilst also providing a challenge to obtain better performance when applying the decision tree algorithm to compare patterns identified by the authors cited in Table 3.

Method

Figure 2 shows the methodology applied to identify factors that influence university dropout. The description of each process is detailed below:

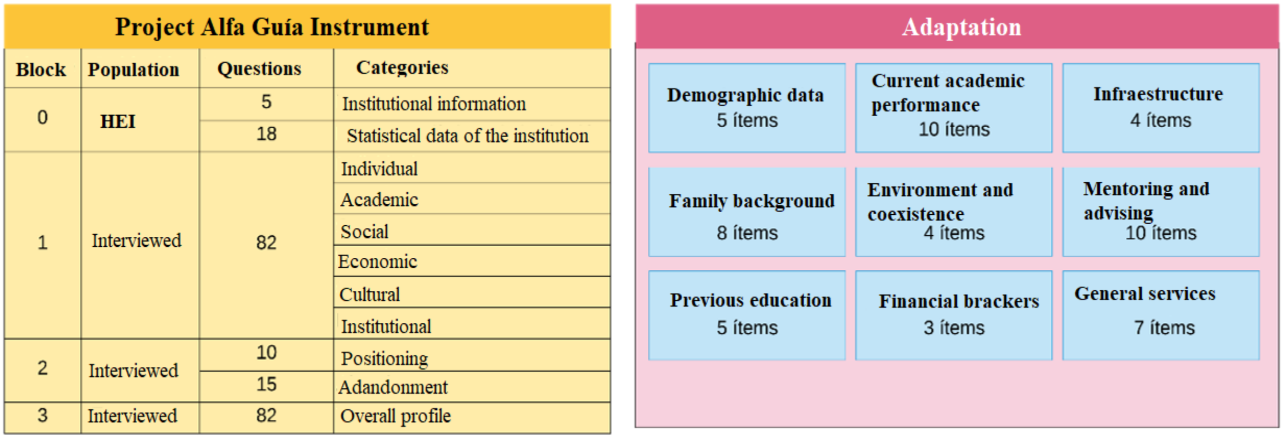
Figure 2. Methodology for determining the most relevant characteristics in university dropouts



Source: Prepared by the authors

The instrument used for data collection was adapted from the International Survey on Dropout in Higher Education, developed by the Alfa Guía project in 2014 (Valle, Eslava, Manzano & García, 2014). The description of this adaptation is shown in Figure 3.

Figure 3. Adapting the instrument for data collection



Project Alfa Guía Instrument

Block	Population	Questions	Categories
0	HEI	5	Institutional information
		18	Statistical data of the institution
1	Interviewed	82	Individual
			Academic
			Social
			Economic
			Cultural
			Institutional
2	Interviewed	10	Positioning
		15	Adandonment
3	Interviewed	82	Overall profile

Adaptation

Demographic data 5 ítems	Current academic performance 10 ítems	Infraestructure 4 ítems
Family background 8 ítems	Environment and coexistence 4 ítems	Mentoring and advising 10 ítems
Previous education 5 ítems	Financial brackers 3 ítems	General services 7 ítems

Source: Prepared by the authors

As can be seen in Figure 3, the adapted instrument has 56 questions divided into 9 categories. These categories are described below:

1. Demographic data: Collects data on age, gender, marital status, number of children and employment.
2. Family history: seeks to identify the educational level of the parent/guardian and oldest sibling, in addition to financial dependency
3. Previous schooling: secondary/high school information.
4. Current academic performance: type of institution, area, average, etc.
5. Financial support: Scholarships, aid, educational credit etc and relevant requirements.
6. Environment and coexistence: Collects information related to the institutional environment.
7. Infrastructure: Estimates the degree of satisfaction with university spaces.
8. Follow-up, tutoring and advice: Identifies the degree of satisfaction with the academic follow-up provided by the academic and administrative part of the institution.
9. Services: Measures the degree of satisfaction with general services provided by the institution to the student.

Although the instrument was adapted, it was considered to have valid reliability according to the Cronbach coefficient (Longest, 2019). In this sense, the instrument had a coefficient of 0.8767 and each item obtained a coefficient of between 0.8 and 0.9, thus reflecting adequate reliability.

Population and sample

The overall population was determined by considering that there are 230,788 higher education students in the State of Puebla, Mexico, according to data from the National Education Statistical Information System [SNIE] (2019). The sample was calculated with a confidence level of 97.5% and an admissible sampling error of 5%. Formula 1 of simple random sampling according to the proportion of finite populations, proposed by Morillas (2014) was used.

Formula 1

$$n = \frac{N\, Z_{1-\propto/2}^{2}\, pq}{(N-1)\varepsilon^{2} + Z_{1-\propto/2}^{2}\, pq}$$

Where: n = sample size; N = population size; Z = confidence level; p = probability of success or expected ratio, (0.5 when value is unknown); q = probability of failure; ε = allowable error. Thus, the value of the sample is observed in Equation 1, with the outcome of 500 respondents being truncated.

Equation 1

$$n = \frac{(230{,}788)(2.24)^2(0.5)(0.5)}{(230{,}788-1)(0.05)^2+(2.24)^2(0.5)(0.5)} = 500.7$$

Simple random sampling was used of several public and private universities in the City of Puebla. Participants were requested to complete the instrument online. In order to obtain the total number of samples required, permission was requested from the directors of various universities located in the City of Puebla and surrounding cities. Where possible, written permission was requested from universities (regardless of whether they were public or private) and groups of students who were taking classes in a computer laboratory (regardless of the degree program and semester in which they were enrolled). In this way, 300 students from public HEIs and 200 students from private HEIs participated. Of these, 311 are female (176 from public HEIs and 135 from private HEIs) and 188 are male (124 from public HEIs and 65 from private HEIs).

Instrument application. It should be mentioned that a form was designed using Google forms (Google, 2019). This enabled the instrument to be administered in a way that was easier to collect information, thus, facilitating the knowledge discovery process (the form can be accessing following the link https://bit.ly/3dVKgJy).

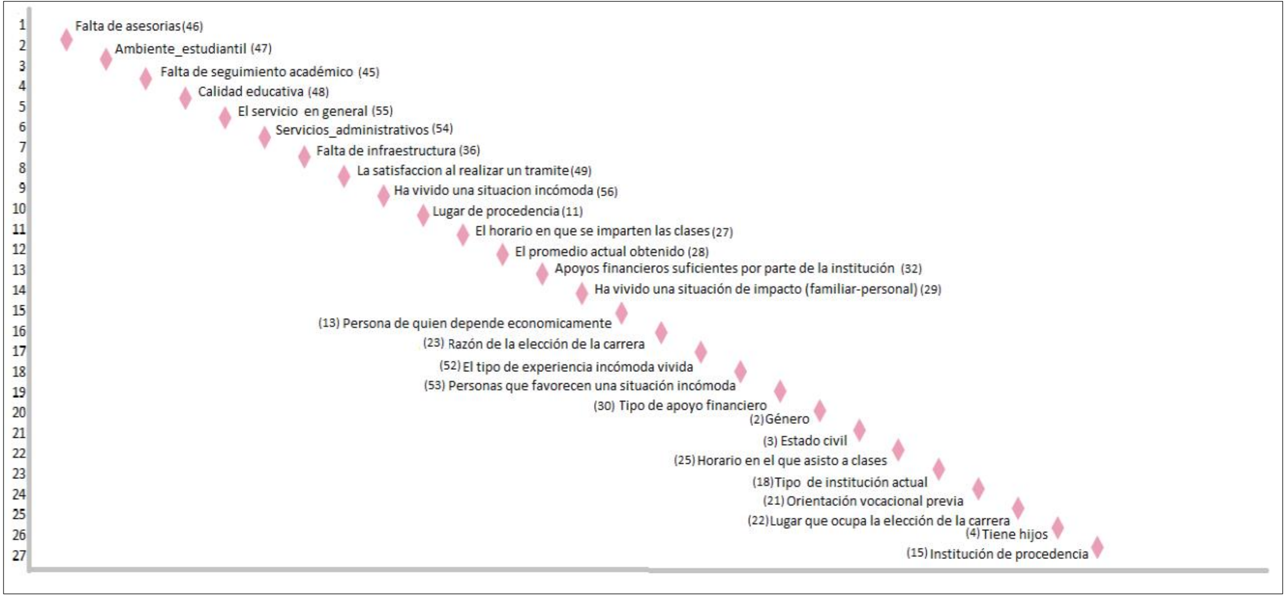
Results

This section presents the application of the knowledge discovery process in two phases: 1) Identification of the most relevant factors by applying attribute selection evaluators, and; 2) Recognition of university dropout patterns via decision trees.

Most relevant factors affecting university dropout

As described above, one of the phases of the knowledge discovery process is the selection of attributes. Figure 4 presents the results obtained following application of the two chosen evaluators (GainRatioAttributeEval and InfoGainAttributeEval), denoting a list of the 27 main attributes from the dataset and containing 56 attributes. It is noticeable that the ranking produced when using both evaluators considers the same attributes to be the most important, despite them having different weights. Given that the results of both evaluators coincide, they are considered to form the basis of a set of rules to favour the identification of early dropouts using decision trees.

Figure 4. The most important attributes in university dropouts using attribute selection

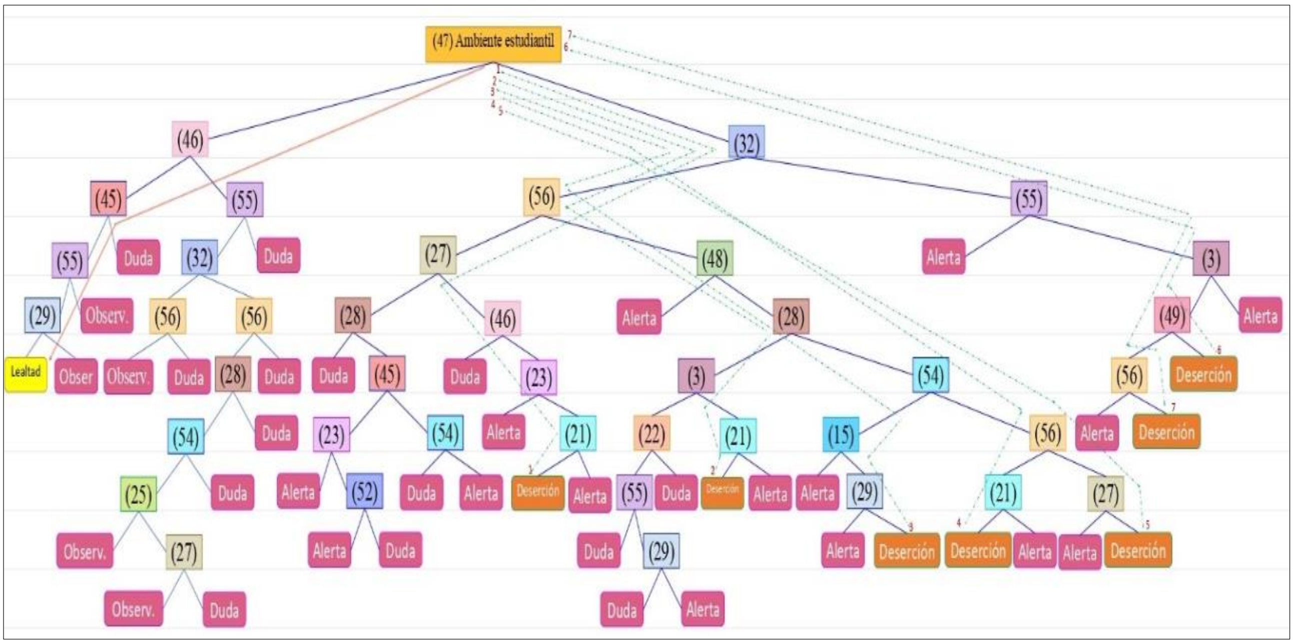


Source: Own elaboration

Patterns of impending defection

In this phase, findings following application of the C4.5 decision tree algorithm in its two phases are presented. In the first phase the complete dataset was used, that is, all 56 attributes were included, obtaining a 74.6% accuracy performance (Table 6). In the second phase, the dataset was modified based on the application of attribute selection as described above. In other words, the C4.5 decision tree algorithm was again applied to the dataset but with only the 27 most relevant attributes being included according to the attribute selection algorithm. This obtained 92.6% performance of accuracy.

Figure 5. Decision tree with the main classified causes predicting imminent university drop-out



Source: Own elaboration

Figure 5 shows the tree obtained using the main identified attributes and the rules used to define an explanatory pattern of young people committed to staying at university. This pattern is based on the following rule:

```
47_STUDENT ENVIRONMENT <= 3
|  46_ LACK OF ADVICE<= 2
|  |  45_ LACK OF ACADEMIC FOLLOW-UP<= 2
|  |  |  55_GENERAL SERVICE <= 1
|  |  |  |  29_ HAS EXPERIENCED A SITUATION OF IMPACT<= 1: LOYALTY
```

The rule reads as follows: IF the student environment is satisfactory to neutral, the lack of counselling and academic follow-up is almost non-existent, the service in general is satisfactory and I have not experienced an impact situation (death of a relative, divorce of parents, breakup) THEN student remains at university. In other words, a university student decides to continue with their studies if there is a totally satisfactory or satisfactory student environment, advise, academic follow-up and services in general are totally satisfactory or satisfactory, and they have not experienced an impact situation. On the contrary, seven rules were found to identify imminent dropout (Figure 5). Examples are given through rule 1 and rule 2, and are described below

Rule 1

```
47_STUDENT ENVIRONMENT > 3
|  32_ INSUFFICIENT FINANCIAL SUPPORT<= 4
|  |  56_ HAS EXPERIENCED AN UNCOMFORTABLE SITUATION<= 3
|  |  |  27_ CLASS SCHEDULE>4
|  |  |  |  46_ LACK OF ADVICE >2
|  |  |  |  | 23_REASON FOR CAREER CHOICE>2
|  |  |  |  |  |  21_RECEIVED PREVIOUS GUIDANCE<=1:DROPOUT
```

Rule 2

```
47_STUDENT ENVIRONMENT > 3
|  32_ INSUFFICIENT FINANCIAL SUPPORT <= 4
|  |  56_ HAS EXPERIENCED AN UNCOMFORTABLE SITUATION > 3
|  |  |  48_GENERAL SERVICE >2
|  |  |  |  28_ CURRENT AVERAGE >4
|  |  |  |  | 54_ADMINISTRATIVE SERVICES <=4
|  |  |  |  |  |  15_INSTITUTION OF ORIGIN = PUBLIC
|  |  |  |  |  |  |  29_HAS SEEN AN IMPACT SITUATION >2: DROPOUT
```

Rule 2 can be explained as follows: IF the student environment is neutral to totally unsatisfactory, financial support provided by the institution is insufficient, student has experienced an uncomfortable situation (harassment, discrimination, mistreatment), the general service and average marks are unsatisfactory, administrative services are totally satisfactory or satisfactory, and student comes from a public high school and has experienced an impact situation THEN student decides to dropout.

In order to know the performance of the decision tree algorithm and classify the "prognosis" class identified according to 1: Loyalty, 2: Observance, 3: Doubt, 4: Alert, and 5: Dropout, two classification matrices were obtained following their examination with the dataset using: a) all 56 attributes and, b) the 27 most relevant attributes (Table 5).

Table 5. Confusion matrix recovered using a) all 56 attributes and b) the 27 most relevant attributes.

a) Classification with 56 attributes					
a	**b**	**c**	**d**	**e**	**Classification**
65	0	24	0	4	a=Compliance
3	102	29	13	0	b=Alert
21	30	115	3	0	c=Doubt
1	19	0	42	0	d=Dropout
3	0	0	0	4	e=Loyalty

b) Classification with 27 attributes					
a	**b**	**c**	**d**	**e**	**Classification**
88	0	4	0	1	a=Compliance
1	135	10	1	0	b=Alert
4	6	179	0	0	c=Doubt
0	8	1	53	0	d=Dropout
1	0	0	0	8	e=Loyalty

Source: Own elaboration

In Table 5, values were selected which were located in the diagonal of each of the matrices, indicating correctly classified instances according to forecast type (classification). For example, if we want to make a comparison between the matrices related to the attribute <Dropout>, we observe that in matrix a, 42 instances were correctly classified as dropout, whilst 1 was incorrectly classified as observance and 19 as alert. In comparison with matrix b, we see that 53 instances were correctly classified as dropout, 8 were correctly classified as alert and 1 as doubt. Thus, in matrix b, the number of correctly classified instances was higher than in matrix a. This occurrence is also observed for the remaining classifications. Likewise, the metrics used to estimate algorithm performance are presented in Table 6.

Table 6. Decision tree performance metrics applied in two phases

Classification	Accuracy	Precision	Recall	F-Measure	Correctly sorted instances
56 attributes	74.6	74.9	74.6	74.6	373
27 attributes	92.6	92.7	92.6	92.6	463

It is observed that using the attribute selection algorithm prior to the C4.5 algorithm significantly increases the accuracy and precision of the algorithm, without losing sight of the fact that the most important outcome is to uncover the most important factors predicting dropout decisions. It is of equal importance to identify the rules or patterns employed by university students as they move towards the desertion of studies.

Discussion and conclusions

As a result of the application of educational data mining, it is possible to conclude that the five main causes of university dropouts are: lack of counselling, inadequate student environment, lack of academic follow-up, poor educational quality and poor service in general. These findings contrast sharply with those found in the literature where the main factor is course dissatisfaction, followed by lack of economic resources, change of marital status, distance between home and the study centre, family reasons, and other causes.

Also, through the use of different evaluators to select attributes it was possible to identify that aspects related a lack of contentment with one's course are found to be between 13th and 19th in the list of the 27 main factors. Further, factors related to economic aspects occupy the 17th and 25th positions in the ranking. This indicates that young university students now report a wider array of factors that are relevant for satisfying their desire continue studying. Thus, in the case of the present sample, relevant factors differ from those extracted from the studies analysed in Table 2.

On the other hand, thanks to the application of the decision tree algorithm, it was possible to establish a series of 7 patterns that lead to school dropout. Likewise, it was possible to classify patterns that are worthy of attention as shown in Figure 5. These include patterns which evidence the presence of doubt, observance or warning. Likewise, the pattern characterising university students' decisions to stay at the institution has been determined. These students perceive an adequate student environment, appropriate academic assessment and follow-up provision and adequate services in general, and do not experience damaging impact situations.

It is known that the majority of HEIs have implemented tutorial and advisory actions, however, it seems that teacher/tutors are assigned an excessive number of students, preventing the service from being efficient or rendering it non-existent. The findings obtained in the present study call for application of HEI guidelines to create mechanisms which support counselling services. This should occur through actions which help to offer better services, such as engagement of final year students with offering guidance to first-year students. In the same way, teachers should offer individualised tutoring to address academic issues and detect deficiency or risk, channelling students to appropriate services.

Further, in order to promote a positive educational climate through university community approaches, parents should be engaged. Coexistence is a dominant factor and should be targeted in order to increase communication, respect, emotional intelligence, conflict resolution and person-to-person approaches, in order to achieve sensitivity, motivation and empathy. At the same time, strengthening the quality of education is crucial in order to ensure that students acquire the competencies and skills required to continue higher studies and successfully enter the workplace.

Finally, it has been seen that dropout does not depend on a single factor. Instead, it is caused by a set of factors and the way in which they interact. This was clearly observed in the patterns recovered through the C4.5 decision tree. It is possible to deduce that such patterns may vary even in similar contexts due to the region, place of origin, socio-economic level, or even individual beliefs, as was also reviewed in the literature (Table 2).

Thus, this study provides an introduction to educational data mining in order to find patterns that prevent imminent dropout and enable preventative action. In order to enrich findings, it is imperative to expand the sample to other states or cities so that various algorithms can be applied to provide more information. This will lead to the establishment of accurate mechanisms to decrease the increasing university dropout rates being reported year after year. This is crucial because despite the fact that federal or state governments commit towards the implementation of mechanisms, these have not been sufficient to decrease these rates with evidence showing that dropout has instead increased in recent years (2016-2019).

References

Abarca R., A., & Sánchez V., M. A. (2005). La deserción estudiantil en la educación superior: el caso de la Universidad de Costa Rica. *Revista Electrónica "Actualidades Investigativas en Educación", 5*, 1-22. https://bit.ly/35TVeLE

Abu-Oda, G. S., & El-Halees, A. M. (2015). Data mining in higher education: university student dropout case study. International Journal of Data Mining y Knowledge Management Process (IJDKP), 5(1), 15-27. https://doi.org/10.5121/ijdkp.2015.5102

Agaoglu, M. (2016). Predicting instructor performance using data mining techniques in higher education. IEEE Access, 4. https://doi.org/10.1109/ACCESS.2016.2568756

Al-Barrak, M. A., & Al-Razgan, M. (2016). Predicting student's final GPA using decision trees: a case study. International Journal of Information and Education Technology, 6(7), 528-533. https://doi.org/10.7763/IJIET.2016.V6.745

Alkhasawneh, R., & Hargraves, R. H. (2014). Developing a Hybrid Model to Predict Student First Year Retention in STEM Disciplines Using Machine Learning Techniques. *Journal of STEM Education: Innovations and Research*, 5(3), 35-42. ERIC. https://bit.ly/2Rd04hi

Aulck, L., Velagapudi, N., Blumenstock, J., & West, J. (2017). Predicting Student Dropout in Higher Education. *Machine Learning in Social Good Applications*, 16-20. https://bit.ly/3aRtae6

Barbosa M. L. M., Serra da Cruz, S. M., & Zimbrão, G. (2014). The Impact of High Dropout Rates in a Large Public Brazilian University: A Quantitative Approach Using Educational Data Mining. *6th International Conference on Computer Supported Education* (págs. 124-129). Barcelona, Spain: INSTICC. https://bit.ly/2ZsYFbD

Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python. USA: O´Really Media, Inc.

Bishop, C. M. (2007). Pattern recognition and Machine Learning. Singapore: Springer.

Cabrera, L., Bethencour, J. T., Álvarez P. P., & González A. M. (2006). El problema del abandono de los estudios universitarios. *RELIEVE, 12*(2), 171-203. https://doi.org/10.7203/relieve.12.2.4226

Carvajal O. P., & Trejos C. Á. A. (2016). Revisión de estudios sobre deserción estudiantil en educación superior en Latinoamérica bajo la perspectiva de Pierre Bourdieu. *Congresos CLABES*. Quito, Ecuador: Escuela Politécnica Nacional. https://bit.ly/2UP9mlT

Cha, G.-W., Kim, Y.-C., Moon, H. J., & Hong, W.-H. (2017). New approach for forecasting demolition waste generation using chisquared automatic interaction detection (CHAID) method. Journal of Cleaner Production, 168, 375-385. https://doi.org/10.1016/j.jclepro.2017.09.025

Chen, W., Xie, X., Peng, J., Wang, J., Duan, Z., & Hong, H. (2017). GIS-based landslide susceptibility modelling: a comparative assessment of kernel logistic regression, Naïve-Bayes tree, and alternating decision tree models. Geomatics, Natural Hazards and Risk, 8(2), 950-973. https://doi.org/10.1080/19475705.2017.1289250

Chiheb, F., Boumahdi, F., Bouarfa, H., & Boukraa, D. (2017). Predicting students' performance using decision trees: Case of an Algerian University. 2017 International Conference on Mathematics and Information Technology (ICMIT). Adrar, Algeria: IEEE. https://doi.org/10.1109/MATHIT.2017.8259704

Dekker, G. W., Pechenizki, M., & Vleeshouwers, J. M. (2009). Predicting Students Drop Out: A Case Study. *2nd International Conference on Educational Data Mining* (págs. 41-50). Cordoba, Spain: International Educational Data Mining Society. https://bit.ly/2ZlH1a3

Delen, D. (2011). Predicting Student Attrition with Data Mining Methods. *Journal of College Student Retention: Research, Theory y Practice, 13*(1), 17-35. https://doi.org/10.2190/CS.13.1.b

Del Pobil, A. P., Mira, J., & Ali, M. (1998). Tasks and Methods in Applied Artificial Intelligence. 11th International Conference on Industrial and Engineering Applications of Artificial In telligence and Expert Systems. 1416. Castellón, España: Springer.

Estrada-Danell, R. I., Zamarripa-Franco, R. A., Zúñiga-Garay, P. G., & Martínez-Trejo, I. (2016). Aportaciones desde la minería de datos al proceso de captación de matrícula de instituciones de educación superior particulares. Revista Electrónica Educare, 20(3), 1-21. https://doi.org/10.15359/ree.20-3.11

Fozdar, B. I., Kumar, L. S., & Kannan, S. (2006). A Survey of a Study on the Reasons Responsible for Student Dropout from the Bachelor of Science Programme at Indira Gandhi National Open University. *International Review of Research in Open and Distance Learning, 7*(3), 1-15. https://doi.org/10.19173/irrodl.v7i3.291

Frank, E., Hall, Mark A., & Witten I. H. (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.

Freitas, A. A. (2002). Data Mining and Knowledge Discovery with Evolutionary Algorithms. The Netherlands: Springer-Verlag. https://doi.org/10.1007/978-3-662-04923-5

Guevara, C., Sanchez-Gordon, S., Arias-Flores, H., Varela-Aldás, J., Castillo-Salazar, D., Borja, M., . . . Yandún-Velasteguí, M. (2019). Detection of Student Behavior Profiles Applying Neural Networks and Decision Trees. 1026, págs. 591-597. Munich, Germany: Springer, Cham. https://doi.org/10.1007/978-3-030-27928-8_9

Gupta, B., Rawat, A., Jain, A., Arora, A., & Dhami, N. (2017). Analysis of Various Decision Tree Algorithms for Classification in Data Mining. International Journal of Computer Applications, 163(8), 15-19. https://doi.org/10.5120/ijca2017913660

Han, J., Kamber, M., & Pei, J. (2011). Data mining: concepts and techniques. Amsterdam: Morgan Kaufmann.

INEGI. (2018). Estadísticas a propósito del día mundial de la población (11 de julio). Ciudad de México: INEGI. https://bit.ly/2xbnZHd

Jadhav, S. D., & Channe, H. P. (2016). Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques. International Journal of Science and Research (IJSR), 5(1), 1842-1845. https://doi.org/10.21275/v5i1.NOV153131

Kabra, R. R., & Bichkar, R. S. (2011). Performance Prediction of Engineering Students using Decision Trees. *International Journal of Computer Applications, 36*(11), 9-12. https://bit.ly/2JdxckV

Kotsiantis, S., Pierrakeas, C., & Pintelas, P. E. (2003). Preventing Student Dropout in Distance Learning Using Machine Learning Techniques. *Knowledge-Based Intelligent Information and Engineering Systems, 7th International Conference* (págs. 267-274). Oxford, UK.: Springer-Verlag Berlin Heidelberg. https://doi.org/10.1007/978-3-540-45226-3_37

Kumar Y. S., Bharadwaj, B., & Pal, S. (2012). Mining Education Data to Predict Student's Retention: A comparative Study. *International Journal of Computer Science and Information Security*, *10*(2), 113-117. https://bit.ly/2JJw9t1

Lavado, P., & Gallegos, J. (2005). La dinámica de la deserción escolar en el Perú: un enfoque usando modelos de duración. Lima, Perú: Universidad del Pacífico. https://bit.ly/39PH3rJ

Londoño A. L. F. (2013). Factores de riesgo presentes en la deserción estudiantil en la Corporación Universitaria Lasallista. *Revista Virtual Universidad Católica del Norte* (38), 183-194. https://bit.ly/1OnjEwM

Longest, K. C. (2019). Using Stata for Quantitative Analysis. California, USA: SAGE Publications.

M.P. van der Aalst, W. (2011). Process Mining: Discovery, Conformance and Enhancement of Business Processes (Google eBook). London, UK: Springer-Verlag. https://doi.org/10.1007/978-3-642-19345-3

Márquez-Vera C., Romero M. C., & Ventura S. S. (2012). Predicción del Fracaso Escolar mediante Técnicas de Minería de Datos. *Revista Iberoamericana de Tecnologías del Aprendizaje*, *7*(3), 109-117. https://bit.ly/2zoZKmo

Márquez-Vera, C., Cano, A., Romero, C., Mohammad N. A. Y., Fardoun, H. M., & Ventura, S. (2016). Early dropout prediction using data mining: a case study with high school students. *Expert Systems*, *33*(1), 107-125. https://doi.org/10.1111/exsy.12135

Mitchell, T. M. (1997). Machine Learning. Singapore: McGraw-Hill.

Mitchell, T. M. (2000). *Decision Tree Learning.* Washington State University. https://bit.ly/2N1AI32

Morales C. J., & Parraga-Alava, J. (2018). How Predicting The Academic Success of Students of the ESPAM MFL?: A Preliminary Decision Trees Based Study. 2018 IEEE Third Ecuador Technical Chapters Meeting (ETCM). Cuenca, Ecuador: IEEE. https://doi.org/10.1109/ETCM.2018.8580296

Morillas, A. (2014). Muestreo en poblaciones finitas. Notas del curso. Málaga-España: Universidad de Málaga. https://bit.ly/2JLLA3K

OECD. (2017). *Skills Strategy Diagnostic Report: Mexico 2017, OECD Skills Studies*. París: OECD Publishing. https://doi.org/10.1787/9789264287679-en

OECD. (2019). *Higher Education in Mexico: Labour Market Relevance and Outcomes, Higher*. París: OECD Publishing. https://doi.org/10.1787/9789264309432-en

Pal, S. (2012). Mining Educational Data Using Classification to Decrease Dropout Rate of Students. *International Journal of Multidisciplinary Sciences and Engineering*, *3*(5), 35-39. https://bit.ly/2xVhAjc

Raju, D. y Schumacker, R. (2015). Exploring Student Characteristics of Retention that Lead to Graduation in Higher Education Using Data Mining Models. Journal of college student retention: Research, Theory y Practice, 16(5), 563-591. https://doi.org/10.2190/CS.16.4.e

Rodríguez-Maya, N. E., Lara-Álvarez, C., May-Tzuc, O., & Suárez-Carranza, B. A. (2017). Modeling Students' Dropout in Mexican Universities. *Research in Computing Science*, *139*, 163-175. https://doi.org/10.13053/rcs-139-1-13

Ruíz C., L. (2009). Deserción en la educación superior recinto Las Minas. Período 2001-2007. *Ciencia e Interculturalidad*, *4*(2), 30-46. https://doi.org/10.5377/rci.v4i1.288

Sara, N. B., Halland, R., Igel, C., & Alstrup, S. (2015). High-School Dropout Prediction

Using Machine Learning: A Danish Large-scale Study. *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning* (págs. 319-324). Bruges, Belgium: i6doc.com. https://bit.ly/2MFgzkp

Secretaría de Educación Pública. (2019). Abandono escolar. Ciudad de México: SEP.

Secretaría de Educación Pública. (2019). Principales cifras del sistema educativo nacional 2018-2019. Ciudad de México: Dirección General de Planeación, Programación y Estadística. https://bit.ly/2yCwivX

Sharma, H., & Kumar, S. (2016). A Survey on Decision Tree Algorithms of Classification in Data Mining. International Journal of Science and Research (IJSR), 5(4), 2094-2097. https://doi.org/10.21275/v5i4.NOV162954

Sivakumar, S., Venkataraman, S., & Selvaraj, R. (2016). Predictive Modeling of Student Dropout Indicators in Educational Data Mining using Improved Decision Tree. *Indian Journal of Science and Technology, 9*(4), 1-5. https://doi.org/10.17485/ijst/2016/v9i4/87032

Universidad Tecnológica de Tabasco. (2019). Glosario de Términos. Villermosa, Tabasco: Universidad Tecnológica de Tabasco. https://bit.ly/2xZ60DK

Ustebay, S., Turgut, Z., & Ali A. M. (2018). Intrusion Detection System with Recursive Feature Elimination by Using Random Forest and Deep Learning Classifier. 2018 International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism (IBIGDELFT) (págs. 71-76). Ankara, Turkey: IEEE. https://doi.org/10.1109/IBIGDELFT.2018.8625318

Vélez, A., & López, J. D. F. (2004). Estrategias para vencer la deserción universitaria. *Educación y Educadores* (7). 177-203. https://bit.ly/39MgeEJ

Valle G. R., Eslava G. G., Manzano P. A., & García M. M. (2014). Encuesta Internacional sobre el Abandono en la Educación Superior. Unión Europea. https://bit.ly/2p8k2Pk

Vijayalakshmi, M., & Kumar, A. S. (2011). Efficiency of decision trees in predicting student's academic performance. Computer Science y Information Technology, 335-343. https://doi.org/10.5121/csit.2011.1230

Vries, W., León A. P., Romero M. J. F., & Hernández S. I. (2011). ¿Desertores o decepcionados? Distintas causas para abandonar los estudios universitarios. *Revista de la Educación Superior, 40*(160), 29-49. https://bit.ly/1TzOzru

Witten, I. H., & Frank, E. (2005). Data Mining: Practical machine learning tools and techniques. San Francisco, CA: ELSEVIER.

Witten, I. H., Frank, E., & Hall, M.A. (2011). Data mining: Practical machine learning tools and techniques (3a. ed.). Morgan Kaufmann Publishers, Burlington. https://doi.org/10.1016/B978-0-12-374856-0.00001-8

Yamao, E., Saavedra, L. C., Campos P. R., & Huancas H. V. D. (2018). Prediction of academic performance using data mining in first year students of peruvian university. *CAMPUS, XXIII*(26), 151-160. https://doi.org/10.24265/campus.2018.v23n26.05

Yang, S., Lu, O., Huang, A., Huang , J., Ogata, H., & Lin, A. (2017). Predicting Students' Academic Performance Using Multiple Linear Regression and Principal Component Analysis. Journal of Information Processing, 170-176. https://doi.org/10.2197/ipsjjip.26.170

Yukselturk, E., Ozekes, S., & Kılıç T. Y. (2014). Predicting dropout student: an application of data mining methods in an online education program. *European Journal of Open, Distance and e-Learning, 17*(1), 119-133. https://doi.org/10.2478/eurodl-2014-0008

Authors / Autores

Urbina-Nájera, A.B. (abunajera@gmail.com) 0000-0002-3700-7287

Argelia B. Urbina Nájera, belongs to the Mexican National System of Researchers. His lines of research focus on the application educational data mining, and of machine learning, data science and business intelligence in the field of education, health and commercial activities. Has a PhD. in Strategic Planning and Technology Management from the Universidad Popular Autónoma del Estado de Puebla (UPAEP), has a Master of Science degree in Computer Engineering from the Universidad Autónoma de Tlaxcala, has a Master of Science degree in Education from the Instituto de Estudios Universitarios. Currently she works a Full-Time Professor-Researcher at Deanery of Engineerings at UPAEP.

Camino-Hampshire, J.C. (josecarlos.camino@upaep.edu.mx) 0000-0002-4686-494

Accenture (México) & UPAEP-University (Mexico). José Carlos Camino Hampshire has a BSc in Industrial Engineering & Management from UPAEP. He also has a MSc in Logistics and Supply Chain Management and a MSc in Data Science and Business Intelligence from the same institution. Currently he works as Management Consulting Manager at Accenture México within the Supply Chain practice for Products Industry (Consumer Goods and Services, retail, hospitality, automotive, among others).

Cruz Barbosa, R. (rcruz@mixteco.utm.mx) 0000-0002-5494-7027

Mixteca´s Technology University (Mexico). Raúl Cruz-Barbosa received his B.S. and M. Sc. degrees from Autonomous University of Puebla, Mexico. He also has a Ph.D. in Artificial Intelligence from Technical University of Catalonia, Spain. Dr. Cruz-Barbosa is member of the Mexican National Research System. His research interests are related to large scale machine learning, digital image processing, data mining and pattern recognition as well as their application in education, Bioinformatics and computer aided detection and diagnosis.

RELIEVE

Revista **EL**ectrónica de **I**nvestigación y **EV**aluación **E**ducativa

E-Journal of Educational Research, Assessment and Evaluation

[ISSN: 1134-4032]