University student dropout prediction for female students in the Computer Engineering career

Ellen L. Méndez Xavier[1,†], Fabrizio Coscia[2,*,†] and Christian von Lücken [3,†]

[1] *Department of Computer Science Education, Facultad Politécnica – Universidad Nacional de Asunción), San Lorenzo, Paraguay*

[2] *Department of Computer Science Education, Facultad Politécnica – Universidad Nacional de Asunción, San Lorenzo, Paraguay*

[3] *Department of Computer Science Education, Facultad Politécnica – Universidad Nacional de Asunción, San Lorenzo, Paraguay*

Abstract

University student dropout is a complex phenomenon raising concern both academically and socially. This issue affects students from different areas and contexts; it appears as the premature interruption of higher education, and it has meaningful consequences both on the people involved and on society as a whole. This paper uses prediction models to analyze academic and socioeconomic data about students of the career of Computer Engineering of the Facultad Politécnica of the Universidad Nacional de Asunción (FP-UNA) focusing on women. These models show their effectiveness to predict dropout, and therefore can be useful tools for educational management and to develop preventive actions.

Keywords

higher education, women, gender, computing, student dropout

1. Introduction

University student dropout is a complex issue that affects students, educational institutions and national social development as a whole. It is defined as a definite or temporary interruption of university studies by a student, who will not get the corresponding academic degree or diploma.

In 2017, the Worl Bank published a study called "*Turning Point: Higher Education in Latin America*". It mentions that only 50% of higher education students get to finish their career and graduate. It is estimated that students in Latin America and the Caribbean take 36% more time on average to finish their career in comparison to the rest of the world [1].

University dropout in Paraguay is associated to economic and family problems, poor academic performance, low motivation and incorrect career choice, among other factors. Also, most of the students in the country have to combine their studies with some type of labor [2].

Data in [3], show that in average, less women than men enter university, but once they are registered, their rate of career completion is higher than that of men. The same behavior has been seen in a specific study in the context of computer careers in the Facultad Politécnica of

Proceedings XVI Congress of Latin American Women in Computing 2024, August 12–16, 2024, Bahía Blanca, Argentina.

* Corresponding author.

† These authors contributed equally.

emendez@pol.una.py (E. Mendez); fabricoscia@fpuna.edu.py (F. Coscia); clucken@pol.una.py (C. von Lücken);

0000-0001-6063-1769 (E. Mendez); 0009-0002-9380-9900 (F. Coscia); 0000-0002-2198-1237 (C. von Lücken)

the Universidad Nacional de Asunción (FP-UNA), which indicates that the number of women in comparison to that of men is lower at the time of entering the university, they are more effective if we standardize the number to evaluate graduation [4]. Also, in [4], we can see that women are disproportionally underrepresented in these careers.

Even though some common challenges have been identified for both men and women, such as curricular and financial barriers to achieve success, women have their own difficulties due to cultural elements and gender stereotypes that can affect their academic performance. These questions may also include home responsibilities and the lack of a family support system while pursuing their engineering studies [3] [4].

This paper takes into consideration academic performance data, as well as socioeconomic data, in order to evaluate dropout possibilities for women and provide a list of students with high dropout risk. We have used a number of techniques from Data Mining (DM) to evaluate these data, to allow us to automatically find the relationships in the data group analyzed. This process has the potential to improve academic management and, as a consequence, improve student retention. In Section II, we explain the analysis methods applied. In Section III we present the results, and then the main conclusions of this paper.

2. University Dropout.

From an individual point of view, dropping out means failing to complete a determined course of action or failing to reach a wanted goal, in pursuit of which the person entered a higher education institution in particular. Therefore, dropping out depends not only on individual intentions but also on social and intellectual processes through which people create the goals they want to achieve in a given university [11]. In other words, Tinto defines dropping out as a situation a student faces when they aspire to achieve something but cannot manage to finish their educational project.

2.1. Theoretical Dropout Model.

According to Tinto, students act according to the exchange theory for building their social and academic integration, expressed in terms of goals and institutional commitment levels. The researcher defines the dropout theoretical model based on two main components:

- Intention and goals: students' willingness to finish their career.
- Institutional commitment: degree of attachment to or identification with the university they belong to.

The interaction with the academic system and the social system of the university modifies a student's perception in terms of the two components defined previously, and they influence the decision of dropping out of or remaining as part of the institution.

3. Information Processing.

The *Knowledge Discovery in Databases* (KDD) process is a methodology used to discover knowledge from large data groups. This is no trivial process since it allows identifying valid, original, potentially useful and understandable data patterns.

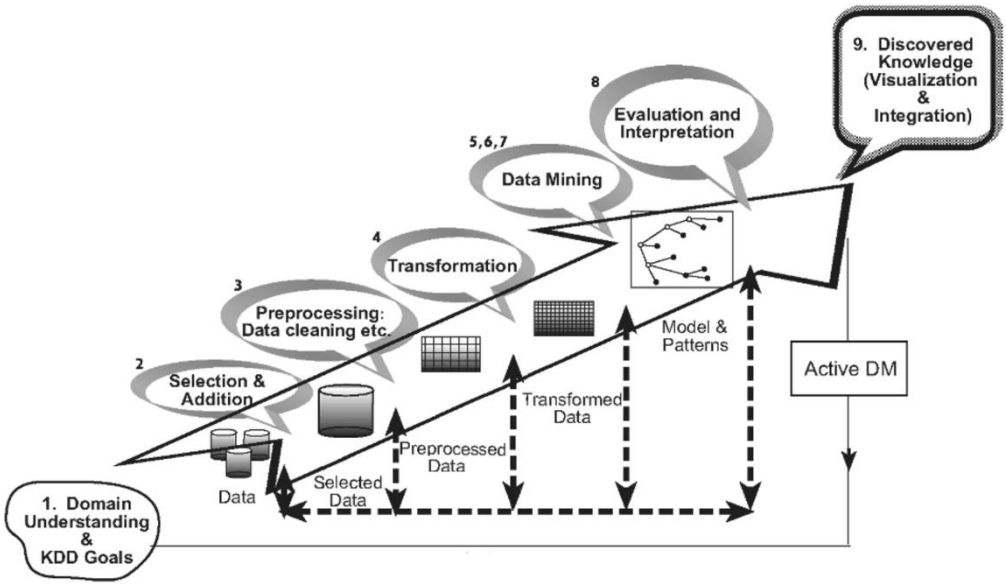

Figure 1: KDD Process [5].

Figure 1 presents the KDD process. It can be seen that Data Mining is a relevant element in this process. Likewise, data preparation, selection and cleaning are no less important, together with the incorporation of previous knowledge and result interpretation. [5]. These steps, when applied in an iterative and interactive manner, facilitate extracting useful knowledge from analyzed data.

This paper has considered the KDD process for data preparation, pre-processing and transformation from an academic database, and the socioeconomic characterization of students of the Computer Engineering career in FP-UNA. The data was used as an entry point to apply different data mining algorithms with the KNIME[2] software. Later, we worked on result fragmentation and comparisons.

3.1. Algorithms.

For the proposed data analysis, the selected algorithms are Decision trees, *Naïve Bayes* and *Random Forest* [8]. These algorithms have proven to be useful to make predictions for data groups that require identifying patterns from different variables such as the university dropout problem requires with academic and socioeconomic data.

They were chosen based on the fact that they are widely known, they have low complexity but are very efficient and also because this is the first time they have been used in our context. Decision trees and *Naive Bayes* are often the first algorithms taught to students due to their simplicity, interpretability and strength in a variety of applications [9].

3.1.1. Decision Tree.

Tree structure, where every internal node indicates a test on an attribute, every branch represents a test result, and every leave node (or terminal node) has a class label [12].

[2] KNIME is a data mining tool that allows developing models in a visual manner. https://www.knime.com/knime-analytics-platform

Simulating a path for a data group instance is as follows: initially we bear in mind we have already generated a tree from training data, also the classification instance has the following characteristics:

- **Average entering score:** 80
- **Has children:** No
- **School modality:** Technical
- **Civil status:** Single

According to the tree branches, the nodes are evaluated, and the corresponding label instance is located.

3.1.2. Naïve Bayes.

The algorithm is based on a conditional probability concept and aims at giving higher importance to those events that are really relevant to the data group [10].

Conditional probability P(A|B) is the probability of event A happening, knowing that another event B is also happening, and it is expressed as the following equation:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

The assumption of attribute independence is generally a poor assumption and is often breached for true data groups [13].

For the simulation of the functioning of this algorithm considering the same instance used for the decision tree, the two main ones are defined first:

- A: The student drops out of the career
- B: The student does not drop out of the career

Then, through training data, the different probabilities for every event are calculated with the characteristics of the evaluated instance:

- P(A) = 0.4
- P(B) = 0.6
- P(Average score = 80 | A) = 0.3
- P(Average score = 80 | B) = 0.4
- P(Has children = No | A) = 0.6
- P(Has children = No | B) = 0.8
- P(School modality = Technical | A) = 0.5
- P(School modality = Technical | B) = 0.6
- P(Civil status = Single | A) = 0.7
- P(Civil status = Single | B) = 0.5

Finally, in order to determine whether the evaluated student is dropping out of the career or not, the product of the probabilities involved in every case is calculated:

$$P(A) = 0.4 * 0.3 * 0.6 * 0.5 * 0.7 = 0.0252$$
$$P(B) = 0.6 * 0.4 * \ 0.8 * 0.6 * 0.5 = 0.0576$$

According to the results, it can be seen that the evaluated student is not dropping out of the career.

3.1.3. Random Forest.

In [10] it is mentioned that there are two approximations in the evolution of data mining:

- Developing new algorithms (which happens very occasionally), or adjusting the parameters of a well-known algorithm.
- Combining classifiers that are kind of simple in order to create a more complex one.
- When the classifiers are decision trees, the resulting algorithm is called *Random Forest*.

In order to simulate the functioning of this algorithm, considering the same instance as in the previous issue, first of all it is assumed that 3 trees are used to form the algorithm and by combining training data, each of them takes a main attribute as root node. Then the instance is evaluated by every tree and each one defines if the career is abandoned or not. Finally, the different decisions are counted and the one with the highest number is the final decision (see Figure 2).

Tree 1:
"Average entering score" as the main characteristic.
Average score > 70 -> No dropping out

Tree 2:
"School modality" and "Civil status" as relevant attributes.
Technical modality and single students have higher dropout probability
-> Dropping out

Tree 3:
"Has children" as a main attribute.
Students without children have lower dropout rates -> No dropping out.

Figure 2: Trees. Example of *Random Forest* rule processing.

Taking into consideration the three trees, the student does not drop out of the career.

3.2. Data Classification.

The data used are divided into three major groups:

1. **Admission data**: Provided by the department of Information Technology and Communication (ITCs) of FP-UNA, including subject scores required for admission to a career in the Polytechnical University. Also, there are other data included, such as gender, home address, telephone number, civil status, city, date and place of birth, among others.
2. **Socio-economic data**: Provided by the department of Information Technology and Communication (ITCs) of FP-UNA, the data are gathered through surveys at different moments of the school year. They include data related to the student´s social and

economic situation. These data include if the student has a job or not, information in relation to the kind of high school training received, family data, among others.

3. **Subject data during the career:** It includes the grades of the 52 subjects of the students of the career.

3.3. Data Characterization.

In the sample there are 1,676 students who entered the university from 2000 until 2023, for the Computer Engineering career, with the gender distribution seen in Figure 2. Since the focus of this paper is analyzing female student dropout, we will be looking at the corresponding data for women (373 students).

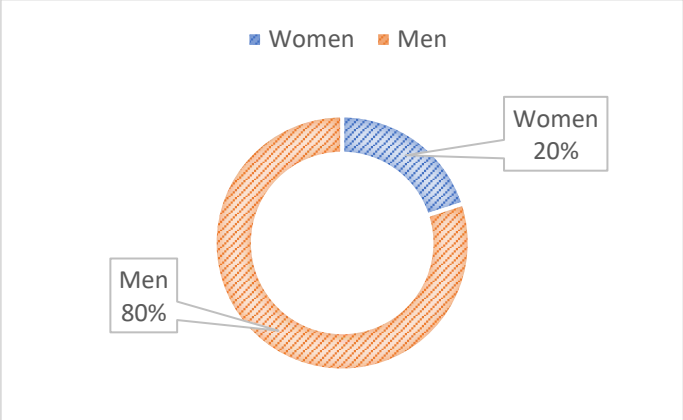

Figure 2: Student distribution by gender.

The career admission process requires passing a number of examinations of basic knowledge for the career and getting the highest scores until taking the number of seats available. The data from these evaluations are used to obtain the admission average. A standardization is made considering the subjects evaluated for admission present yearly variations, thus applying the Admission Average.

For the group of data referring to student socioeconomic information, the following attributes were selected considering the possible values indicated below.

Table 1
Socioeconomic attributes and their possible values

Attributes	Values
Department	17 possible departments in Paraguay
Civil Status	Single – Married - Divorced – Other
Do they work?	Yes – No
Do they have children?	Yes – No
Gender	M – F
School Modality	Technical – Scientific
School type	Public – Private – Subsidized – Cooperative – Other

Finally, for academic data, we looked at all the Computer Engineering career subjects except the Final Dissertation Paper (FDP). Together with the Department of Academic Statistics of FP-

UNA, it was decided that the average of all of a subject's attempts should be taken into account in order to reflect the student's whole academic process.

In order to identify the characteristics of those students who have graduated from the career, the Academic Department provided a subgroup of academic data from 265 Computer Engineering students who have already finished their studies. This way, the GRADUATION attribute was created, and the dropout concept was applied as explained in the following paragraph.

3.4. Data Pre-processing.

Before introducing data in the KNIME tool, they went through preprocessing for information cleaning and optimization, so that it can be used better in the mentioned software.

Initially, there were 6 files with the following characteristics:

- **21042023_nota_IIN.csv:** this includes students' grades and other data about those grades.
- **21042023_nota_retroactiva_IIN.csv:** the same as the previous file but it has students' older grades.
- **28042023_nota_IIN_convalidaciones.csv:** it has validation grades of students who have changed their career.
- **puntaje ingreso.csv:** it has data related to student admission, besides their grades, it also includes sone socioeconomic data.
- **socioeconomica.csv:** it has the answers to the socioeconomic survey carried out by FP-UNA.
- **03042023_egresados_IIN.csv:** it is a list of Computer Engineering students who have finished their career successfully.

The data treatment process was the following:

1. We took the file 21042023_nota_IIN.csv and for every record, we grouped the grades per student, bearing in mind that a student may have more than one grade per subject in case such student failed such subject. Then we repeated the process taking files 21042023_nota_retroactiva_IIN.csv and 28042023_nota_IIN_convalidaciones.csv.
2. Then we processed the file **puntaje_ingreso.csv** by assigning admission scores to the corresponding students. Next, we carried out a similar process with the file **socioeconomica.csv**.
3. The last file, **03042023_egresados_IIN.csv**, was used to add the GRADUATION column and identify those students who have already finished their career.
4. Once all the files were processed, we started to apply the dropping out definition set in order to identify students who are presumed career dropouts.
5. The last step was grouping all the data in a final file that has socioeconomic data, admission average grades, and average grades per subject.

3.5. Dropout Definition at FP-UNA.

For the Universidad Nacional de Asunción there is no formal definition for school dropout. In fact, the bylaws do not mention dropping out at any moment as a matter of discussion or analysis. In order to come closer to a definition of dropping out in FP-UNA, the Coordination of Academic Statistics considers dropping out as a student's lack of registration for four consecutive periods (two years).

3.6. Other considerations.

Data volume in relation to academic information is a lot larger than that of socioeconomic data, also considering that such data are gathered through surveys which some students decide not to answer since they are not mandatory. This makes it necessary to evaluate the impact of combining both groups. Therefore, we applied a strategy of evaluating academic data alone independently and later, academic and socioeconomic data, in order to compare their impact on the application of data mining techniques.

3.7. Model Validation.

The confusion matrix presents a graphic vision of errors made by the classification model in a table. It is a graphic model to visualize the level of correctness of a prediction model. In literature, it is also known as a contingency table or error matrix [10].

	Predicted Class	
True Class	**P**	**N**
P	TP	FN
N	FP	TN

In essence, this matrix indicates the number of instances that have been correctly and incorrectly classified. The parameters indicated are:

- True Positive (TP): number of correct classifications of the positive kind (P).
- True Negative (TN): number of correct classifications of the negative kind (N).
- False Negative (FN): number of incorrect classifications of the positive kind classified as negative.
- False Positive (FP): number of incorrect classifications of the negative kind classified as positive.

From the confusion matrix, we have defined a group of metrics that allow quantifying the goodness of a classification model. An error (ERR) is the sum of incorrect predictions on the total number of predictions. On the contrary, accuracy (ACC) is the number of correct predictions on the total number of predictions, as the following equation shows:

$$ERR = \frac{FP + FN}{FP + FN + TP + TN}$$

$$ACC = \frac{TP + TN}{FP + FN + TP + TN} = 1 - ERR$$

4. Results Obtained

For the application of the different algorithms, the data were divided into two subsamples at random. The first one with 50% of records which were used as training data, and the second one with the other 50% of records which were used to test the prediction ability of the model.

Next, the results of the application of the 3 algorithms with their validation data through a confusion matrix technique are shown, both for academic data processing and for academic and socioeconomic data processing focused on female population.

4.1. Decision Tree

With this algorithm, taking only academic data, we classified 81 instances correctly, while 42 instances were classified incorrectly. This implies a 65.32% correctness rate. In relation to academic and socioeconomic data, the correctness rate is 70%.

The results of the confusion table including the results of academic data analysis and academic and socioeconomic data analysis can be seen in Table 2.

Table 2
Confusion table – decision tree

	Graduation (Prediction)			
	Academic		***Academic + Socioeconomic***	
Graduation (Real)	**NO**	**YES**	**NO**	**YES**
NO	67	14	77	0
YES	29	14	33	0

4.2. Naïve Bayes

We used this algorithm to classify 96 instances correctly, while 28 instances were classified incorrectly. This implies a 77.41% correctness rate for academic data, and a 78.35% rate for socioeconomic and academic data. The results of the confusion table can be seen in Table 3.

Table 3
Confusion Table – *Naïve Bayes*

	Graduation (Prediction)			
	Academic		***Academic + Socioeconomic***	
Graduation (Real)	**NO**	**YES**	**NO**	**YES**
NO	69	12	57	31
YES	16	27	9	27

4.3. Random Forest

With this algorithm, we classified 107 instances correctly, while 17 instances were classified incorrectly. This implies an 89.29% correctness rate for academic data and 84.67% with the inclusion of socioeconomic data. The results of the confusion data can be seen in Table 4.

Table 4
Confusion Table – *Random Forest*

	Graduation (Prediction)			
	Academic		*Academic + Socioeconomic*	
Graduation (Real)	**NO**	**YES**	**NO**	**YES**
NO	69	12	70	18
YES	5	38	1	35

Based on the results obtained in the correctness rate of the algorithms, Figure 3 shows a comparison of the 3 algorithms both for analyzing only academic data and for analyzing academic and socioeconomic data.

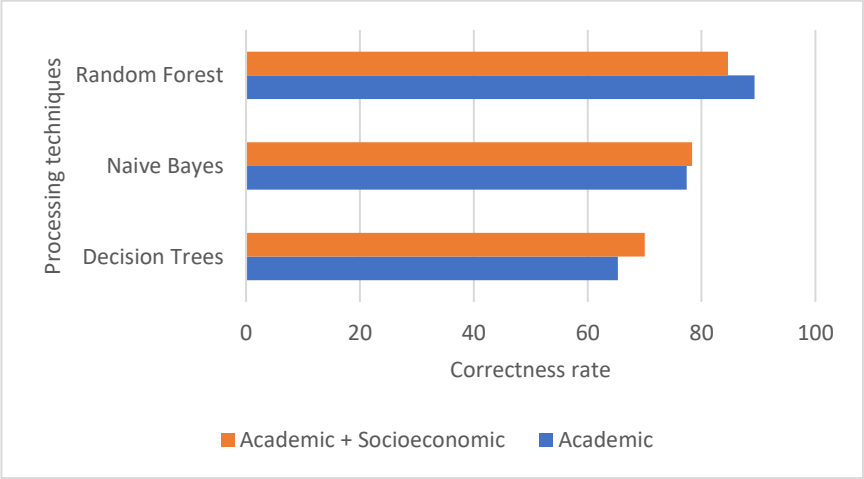

Figure 3: Algorithm efficiency comparison according to data group.

By analyzing the data carefully, the algorithms have identified 123 students with dropout probabilities and have confirmed the validation data. Table 5 shows numbers identified by algorithm in relation to data group.

Table 5

Number of students with dropout probabilities as identified by each algorithm validated with data.

	Academic	**Academic + Socioeconomic**
Decision Trees	67	77
Naive Bayes	69	57
Random Forest	69	70

In the lists provided by the algorithms, there are 19 students who appear in all of them.

The maximum number of coincidences is the appearance of 19 students in all the algorithms executed. In the results of algorithms applied to the academic data group, we see that 52 out of 80 students show up in the 3 lists. In relation to academic data including socioeconomic data, 42 out of 84 students appear in all 3 lists.

5. Conclusions.

In this analysis, we have studied the application of data mining algorithms to predict university dropout, taking into consideration two different stages: one based exclusively on academic data, and another one that includes socioeconomic variables, applied to a data group associated to women exclusively.

Even though all the algorithms were able to identify students who might drop out of the career properly, none of them got the whole group. Also, the group of algorithms was not able to identify all real dropout cases. The existence of a number of students who have been identified by all the algorithms as having dropout probabilities is relevant, other students were identified by 2 of them only. This suggests that students in the result lists could be classified using the number of algorithms as criteria to spot them as students with dropout possibilities, in order to prioritize assistance by the academic manager.

We have seen that we were able to get useful results, but necessary improvements could be achieved by using other algorithms, a higher number of data, different data preprocessing strategies, among others. Even though it is easy to think that a higher number of variables to be analyzed could lead to better predictive capacity, the existing socioeconomic data did not improve results.

For future work, we expect to be able to carry out a detailed characterization analysis of the results obtained in order to identify common patterns and higher repetition of noticeable characteristics. Also, it would be important to carry out an institutional analysis on the current situation of students that have confirmed their dropping out and make a validation of their reasons for doing that.

Special Thanks

To the Coordination of Statistics of the Academic Direction of FP-UNA, for their support during the preparation of this analysis and the approximation to definitions of aspects that have no institutional definition.

References

[1] Mundial, B. (2017). **Momento decisivo: La educación superior en América Latina y el Caribe. Direcciones en Desarrollo**, Washington. *(Decisive moment: Higher education in Latin America and the Caribbean. Developing Guidelinnes, Washington.)*

[2] I. Acuña, "**Acceso y permanencia de estudiantes de educación superior en la ciudad de Pilar**" Ciencia Latina Revista Científica Multidisciplinar, vol. 5, no. 6, pp. 12 372–12 384, 2021. *("Access and stay of higher education students in Pilar City" Multidisciplinary Scientific Magazine Latin Science)*

[3] Madar, N.K., Danoch, A., & Morera, L.S.(2022). **Dropouts of women in engineering studies: A comparable case**. Journal of Entrepreneurship Education, 25(1), 1-10.

[4] Méndez, E. , von Lücken, C., & Cantero, R. (2022). **Applications, admissions and graduations of women in computer science careers for the universidad Nacional de Asunción**. Proceedings http://ceur-ws. org ISSN, 1613, 0073.

[5] O. Maimon and L. Rokach, **Data Mining and Knowledge Discovery Handbook**. Springer, 2005.

[6] P. Ramírez and E. Grandón. **Predicción de la deserción académica en una universidad pública chilena a través de la clasificación basada en Árboles de decisión con parámetros optimizados**, Formación universitaria, vol. 11, no. 3, 2018. *(Academic droput*

prediction in a Chilean public university through classification based on Decision Trees with optimized parameters, University Training, issue 11, number 3, 2018.)

[7] A. Lourens and D. Bleazard. **Applying predictive analytics in identifying students at risk: A case study.** South African Journal of Higher Education, vol. 30, no. 2, pp. 129–142, 2016.

[8] Pranckevičius, T., & Marcinkevičius, V. (2017). **Comparison of Naive Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification**. Balt. J. Mod. Comput., 5.

[9] Witten I., Frank E., & Hall M. (2011). **Data Mining: Practical Machine Learning Tools and Techniques**. Morgan Kaufmann.

[10] Roma, J. C., Quiles, R. C., Roig, J. G., y Alfonso, J. M. (2017). **Minería de datos: Modelos y algoritmos**. Editorial UOC, S.L. *(Data mining: Models and algorithms.)*

[11] Tinto, V. (1989). **Definir la deserción: Una cuestión de perspectiva.** Revista de Educación Superior. 18(71). *(Defining dropout: A matter of Perspective. Higher Education Magazine.)*

[12] Maimon, O. y Rokach, L. (2005). **Data Mining and Knowledge Discovery Handbook**. Springer.

[13] Mosquera, R., Castrillón, O., y Parra, L. (2018). **Máquinas de soporte vectorial, clasificador naive bayes y algoritmos genéticos para la predicción de riesgos psicosociales en docentes de colegios públicos colombianos**. Información tecnológica, 29(6):153–162 *(Vectoral support machines, naive bayes classifier and genetic algorithms for psychosocial risk prediction in teachers from Colombian public schools. Technologic information.)*