Journal of Computer and Communications, 2020, 8, 20-27
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

# Evaluation of Prediction Algorithms in the Student Dropout Problem

**Luis Earving Lee[1], Salvador Ibarra Martínez[1*], José Antonio Castán Rocha[1], Jesús David Terán Villanueva[1], Julio Laria Menchaca[1], Mayra Guadalupe Treviño Berrones[1], Emilio Castán Rocha[2]**

[1]Faculty of Engineering, Autonomous University of Tamaulipas, Tampico, Tamaulipas, Mexico
[2]Tec NM/Madero Institute of Technology, Madero City, Tamaulipas, Mexico

Email: *sibarram@uat.edu.mx

**How to cite this paper:** Lee, L.E., Martínez, S.I., Rocha, J.A.C., Villanueva, J.D.T., Menchaca, J.L., Berrones, M.G.T. and Rocha, E.C. (2020) Evaluation of Prediction Algorithms in the Student Dropout Problem. *Journal of Computer and Communications*, **8**, 20-27.
https://doi.org/10.4236/jcc.2020.83002

**Received:** December 17, 2019
**Accepted:** March 2, 2020
**Published:** March 5, 2020



## Abstract

University dropout is a growing problem which, in recent years, is using computer techniques to assist in the detection process. The paper presents the evaluation of some prediction algorithms to detect a student with a high possibility of scholar desertion. The approach uses real data from past scholar periods to create a dataset with different information of the students (*i.e.*, personal, economic, and academic records). The algorithms selected in the experimental phase were: J48 decision tree, K-near neighbors, and support vector machine. We use two similarity metrics to split the dataset with cases with at least 80% of similarity to evaluate each case. We use the data from 2010 to 2016 with real students' information to predict if there exists the possibility of a real academic dropout in one test for a period. The results show that the J48 algorithm reaches a better performance in both experiments. Besides, the tree generated for each student is taken as a path of attention, reaching around 88% of effectiveness. Finally, the conclusions argue the contributions of the paper and propose a future line of research.

## Keywords

Student Dropout, Prediction Algorithms, J48, K-Near Neighbors, Support Vector Machine

## 1. Introduction

In the last ten years, the scholar desertion arises as one of the most studied problems by academic organisms. A deficient academic level of the population reflects in low productivity in all productive sectors. Unfortunately, it is a problem with a global impact. Recent information provided by the UNESCO [1] in-

dicates that Mexico is the 11 out of 13 countries in Latin America in a total of adult people that achieves a bachelor's degree [2]. Scholar desertion is a phenomenon with a variable behavior that independence of the academic level. In response, the authorities in charge of the educational sector implement strategies to reduce the dropout index of current students. For example, activities between students and the school look for knowing on time the situation of each student whit information provided by the academic department in-charge (*i.e.*, tutorial) [3].

In some cases, the data is not adequate and updated, which reflects errors or actions performed out of time when the student is already deserting. In this sense, the traditional methods have demonstrated not be suitable for avoiding scholar desertions because of data inconsistency or an unsuitable action plan. With the use of a computational paradigm of resolution these problems appear to be tackled.

In reference [4], the authors propose the implementation of a decision tree using the Gini index, as a classification model to detect dropout with students in online courses of the second year. The motivation of the works was to provide information about a possible student dropout at an early stage, so the instructor can react in a feasible way to avoid the desertion. The results demonstrate that a precision around 72% in detecting dropout cases in the $8^{th}$ week of the semester. However, in the same results, only 26% of failing students were discovered in the middle of the period. The authors conclude that it is necessary to increase or evaluate other relevant information to distinguish failing students at any moment of the course.

The reference [5] proposes a multi-objective optimization model based on dynamic data, including classical predictors. The main idea of this work is to find the earliest moment in the student career, in which a reliable prediction is possible. To do that, the paper presents an optimization model tasted with real data without using standard information used by work in the literature (*i.e.*, gender or age). The proposal includes two prediction models. The first one analyses the student situation after two semesters reaching a dropout prediction accuracy in the ranges from 74% to 77%. Meanwhile, the second one evaluates the students after four semesters, with a prediction range between 78% and 83%. The authors suggest the experimentation in other universities with different academic systems to evaluate the accuracy and suitability of their approach.

In [6], the authors argue that academic dropout is a situation that, in some cases, not depends on educational factors. They suggest five main factors criteria as deterministic events to identify a potential dropout such as demography, social interaction, finance, motivation, and personal. The approach process in 3 phases called collecting data, processing data, and modeling the 37 new criteria generated from the five factors. The authors conclude that the prediction of a dropout depends on 2 kinds of combination factor. The first combination is the number of family members and the interest in further study. The second combination factor is the number of family member and the relationship with the lecturers. The author emphasizes that the best model is the second combination

factor with decision tree method using split criterion is maximum deviance reduction and maximum split is 2 with time for training is 1.7386 seconds.

The reference [7] uses knowledge discovery techniques to analyze historical student course grade data to predict a dropout from a particular course. The authors describe a method for predicting student dropout using logistic regression as a classification technique. The implementation of feed-forward neural networks, support vector machine, probabilistic ensemble simplified fuzzy ARTMAP—adaptive resonance theory mapping PESFAM classifier and a system for educational data mining techniques looks for checking the suitability of their proposal. The discussion argues the reduction of dropout rate by 14% concerning previous academic years in which no dropout prevention mechanism was implemented. Despite the promising results reached from medium to long term during the course, the authors indicate that we need to study another more or less automated monitoring technique to evaluate the dropout problem in the early weeks of any class.

All the above works are a feasible example of current approaches with suitable and robust results that looks for providing academic solutions to our students in time and form. However, they only care about the identification of a potential dropout without committing to offer a reliable solution to the lecturer of the course or the academic tutor. In this sense, our approach takes advantage of the solution reached to distinguish among all the factors those three factors that are more incipient when the student decides to abandon not only a particular course but rather a period or more drastic, the complete university.

To follow, Section 2 is presented the main aspects of the proposed approach. Besides, Section 3 is devoted to present the results and their remarks. To the end, in Section 4, conclusions and future work are depicted in order to highlight the advantages of implementing prediction algorithms to detect a student with a high possibility of desertion.

## 2. Implementation of the Approach

The academic sector seeks to implement technology that allows it to be more efficient and appropriate for our times. One example of this effort is the implementation of computational expert systems to support the evaluation entrance psychometric proofs. The present work proposes the generation of a dataset by using a test applied at the moment of the students' entry at each cycle. To put into practice this approach, we use information provided by Engineering School at the Autonomous University of Tamaulipas, Mexico. Such school applies a test using a website.

https://registroaspirantes.uat.edu.mx/Carpeta100/WebFormAsp102.aspx. A dataset with 1,265 students was generated using information from 2012 to 2016. The format of the data response to different types, for example, the variable age, is numerical, and the variable city is a string. We propose to discretize all data to avoid conflicts in the analysis process and transform all data in a qualitative

format (see Figure 1).

When the dataset is completed, the similarity metrics are used to generate a temporal dataset corresponding to the studied case. The metrics implemented are Cosine (1) and Jaccard index (2). Each new student $n_i$ is compared with all the cases in the general dataset $D$ and only those cases with at least 80% of similarity are selected to generate a temporal dataset for such student $D_{ni}$. This process is performed with the idea of reaching a better classification by suppressing cases not near to the studied case.

$$\forall n_i \in N \exists D_{n_i} \in D \mid D_{n_i} \leftarrow D_i \geq \text{metric}\left(n_i, D\right) = 0.8000$$

$$\text{cosine}\left(n_i, D\right) = \frac{\sum_{k=1}^{r} n_{i_k} * D_k}{\sqrt{\sum_{k=1}^{r}\left(n_{i_k}\right)^2 * \sum_{k=1}^{r}\left(D_k\right)^2}} \tag{1}$$

$$\text{Jaccard}\left(n_i, D\right) = \frac{\sum_{k=1}^{r} n_{i_k} * D_K}{\sum_{k=1}^{r}\left(n_i\right)^2 + \sum_{k=1}^{r}\left(D_k\right)^2 - \sum_{k=1}^{r}\left(n_{i_k} * D_k\right)} \tag{2}$$

For illustrative reasons, we perform 1000 experiments to compare each prove a new and different student with the dataset. Figure 2 shows the number

| Real Data | | Transformed Data | |
|---|---|---|---|
| AvePreviousSchool | General Ave | AvePreviousSchool | General Ave |
| 8 | 6.68 | 2 | 3 |
| 0 | 5.46 | 4 | 4 |
| 6.5 | 6.27 | 3 | 3 |

Figure 1. Example of the transformation data.

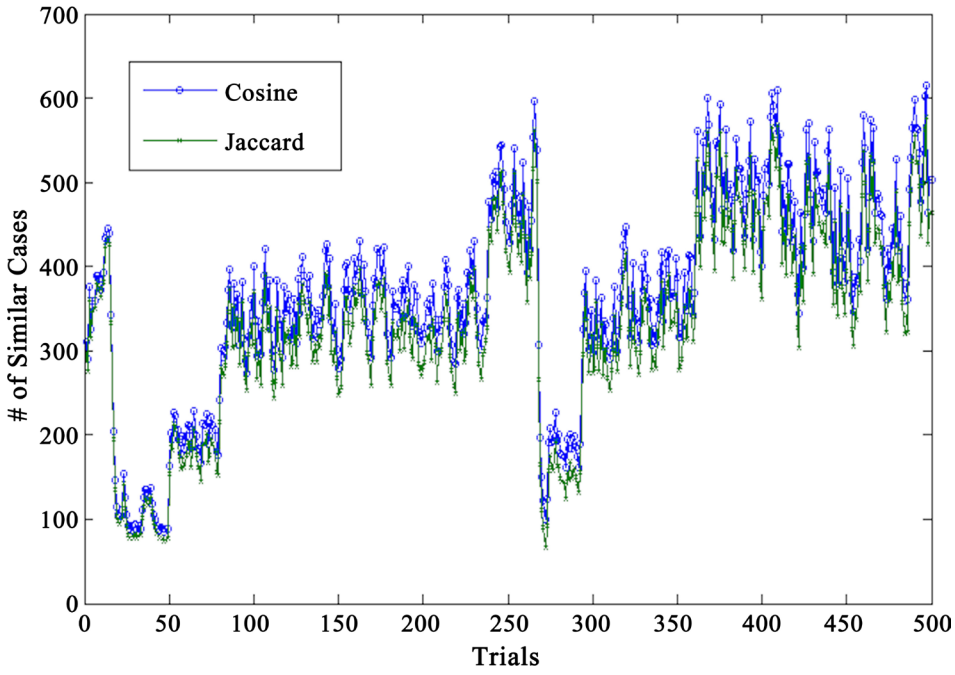


Figure 2. Resume of experiments for recovery cases with similarity metrics.

of recovery cases with at least 80% similarity. The results show that for all trials the cosine metrics reach a great number of recovery cases. The expectative is to take better decision because the temporal dataset only has the most similar cases.

The three prediction algorithms are implemented using WEKA when the dataset is ready. Table 1 presents the classification matrix and Table 2 presents the confusion matrixes to corroborate the performance of each algorithm.

Table 3 resumes the precision and sensitivity of the three algorithms used in this approach. We calculated the measures from the classification matrix presented in Table 1. The accuracy refers to the set of cases classified in the corresponding class, and they are of that class. Sensitivity is the fraction of cases out of all cases that are correctly classified. Sensitivity measures the probability

Table 1. Scheme of classification matrix.

| | | Current Class | |
|---|---|---|---|
| | **Categories** | 0 | 1 |
| **Hypothetical Class** | 0 | TN | FN |
| | 1 | FP | TP |
| **Total Columns** | | **N = FP + TN** | **P = TP + FN** |

Table 2. Confusion matrix for each algorithm.

| **J48 algorithm confusion matrix** | | | | |
|---|---|---|---|---|
| | **0 (dropout) real** | **1 (student) real** | **0 (dropout) real** | **1 (student) real** |
| 1 (dropout) | 474 | 24 | 93.31 | 27.59 |
| 0 (student) | 34 | 63 | 6.69 | 72.41 |
| Correct | 474 | 63 | 93.31 | 72.41 |
| Wrong | 34 | 24 | 6.69 | 27.59 |
| **KNN algorithm confusion matrix.** | | | | |
| | **0 (dropout) real** | **1 (student) real** | **0 (dropout) real** | **1 (student) real** |
| 1 (dropout) | 462 | 36 | 87.17 | 55.38 |
| 0 (student) | 68 | 29 | 12.83 | 44.62 |
| Correct | 462 | 29 | 87.17 | 44.62 |
| Wrong | 68 | 36 | 12.83 | 55.38 |
| **SVM algorithm confusion matrix.** | | | | |
| | **0 (dropout) real** | **1 (student) real** | **0 (dropout) real** | **1 (student) real** |
| 1 (dropout) | 479 | 19 | 87.57 | 39.58 |
| 0 (student) | 68 | 29 | 12.43 | 60.42 |
| Correct | 479 | 29 | 87.57 | 60.42 |
| Wrong | 68 | 19 | 12.43 | 39.58 |

Table 3. Resume of measure obtained by the prelaminar results.

| | J48 | KNN | SVM |
|---|---|---|---|
| Accuracy | 0.952 | 0.928 | 0.962 |
| Sensitivity | 0.933 | 0.872 | 0.876 |
| F-measure | 0.942 | 0.899 | 0.917 |
| RocArea | 0.861 | 0.717 | 0.63 |
| RMSE | 0.277 | 0.401 | 0.382 |
| TPRate | 0.952 | 0.928 | 0.962 |
| FPRate | 0.351 | 0.701 | 0.701 |

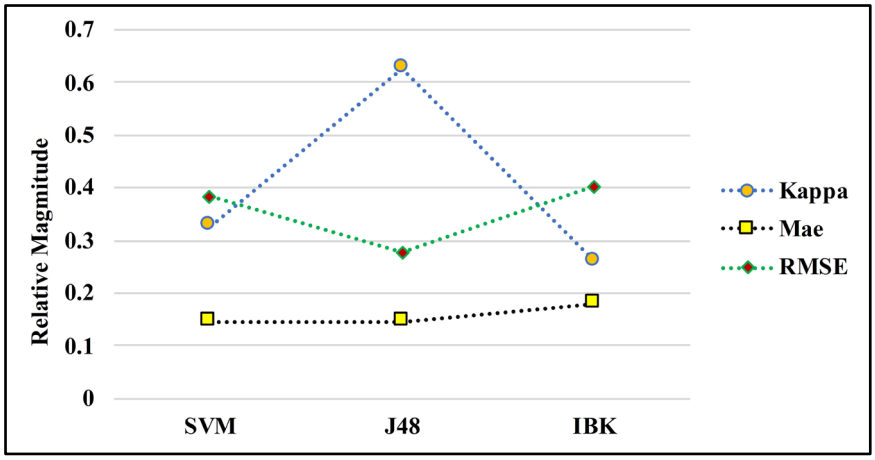


Figure 3. Resume of statistic analysis of the data.

of the algorithm to assign a student of a particular class in such a category. In these results, we see that J48 reaches a better overall performance. Finally, figure 3 shows another interesting statistical analysis of the classification results of the tree proved techniques.

## 3. Experiments and Results

The data from 2012 to 2016, represents 10 periods of classes with a particular student dropout behavior (see Figure 4). Then, we perform 10 experiments (one for each period) to validate the suitability of the proposal approach confronting the J48 technique versus real cases. For example, in the period of classes from January to June 2012 (2012-1) there are 50 dropouts registered but the system only detects 44 cases. The results indicate that J48 is a suitable and useful method to classify all the dropouts. But using the tree of each student as a possible solution to avoid desertion, only 88% of the cases match with the historical information about the reason for the dropout of each student. It is crucial to emphasize that the experiments only compare the highest node in the tree with the academic profile of the student. Figure 4 summarizes the results where each bar represents the group of real dropouts and the line is the amount of reached solution for the J48 technique.

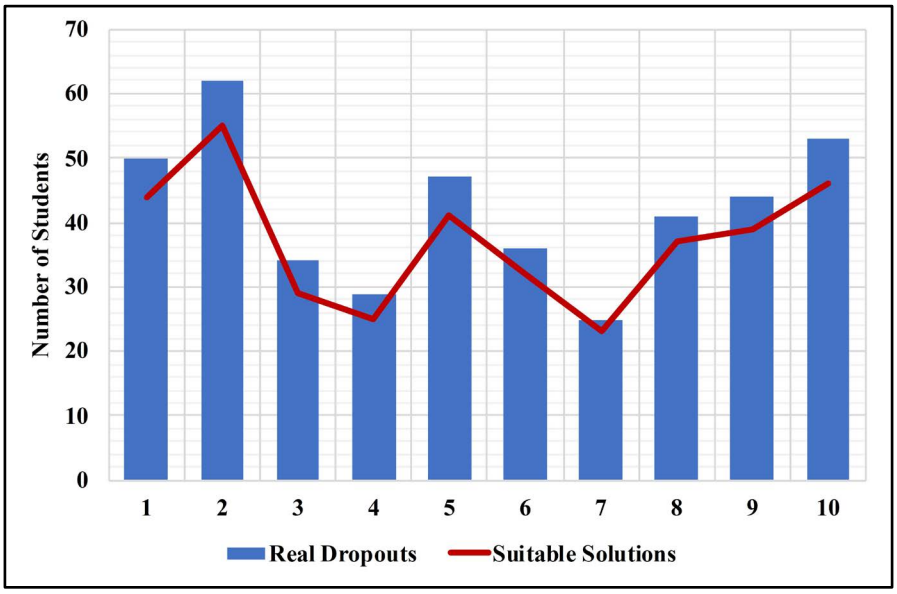


Figure 4. Dropouts in experiments and their match with a suitable solution.

## 4. Conclusion

The methodology proposed aims to be a suitable technological solution to early detection of a scholar desertion. The implementation of a similarity metric ensures the selection of the best cases for the application of the selected predicting algorithms. Although the proves can clearly demonstrate that the two metrics reach a closer, the Cosine metric recoveries more similar cases than the Jaccard index. We observe that more cases mean a better prediction. The paper proves three prediction algorithms with the data recovery for the similar metrics, a decision tree, a support vector machine, and a k-nearest neighbor. These methods were evaluated by statistics values to determine the most effective one. The decision tree reaches a better overall performance for detecting student with high possibilities of deserting. With real data of 100 deserting students, the decision tree demonstrates that it is capable of detecting all cases. We define a personal path of attention for each student, where the tutor must consider each node form the top to the down, and 88% of the cases were detected. However, there is still a long way to go on this exciting and crucial academic topic. For example, it would be essential to evaluate student behavior through the use of multi-objective algorithms that assess the skills and emotions of the person interested in a profession.

## Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

## References

[1] UNESCO (2014) América Latina y el Caribe Revisión Regional 2015 de la educación para Todos. Santiago de Chile.

[2] Gonzalez, A.G.H., Armenta, R.A.M., Rosales, L.A.M., Barrientos, A.G., Xihuitl, J.L.T. and Algredo, I. (2016) Comparative Study of Algorithms to Predict the Desertion in the Students at the ITSM-Mexico. *IEEE Latin America Transactions*, **14**, 4573-4578. https://doi.org/10.1109/TLA.2016.7795831

[3] Alcover, R., Benlloch, J., Blesa, P., Calduch, M.A., Celma, M., Ferri, C., *et al.* (2007) Análisis del rendimiento académico en los estudios de informática de la Universidad Politécnica de Valencia aplicando técnicas de minería de datos. *Pace Pacing and Clinical Electrophysiology*, **164**, 170.

[4] Sukhbaatar, O., Ogata, K. and Usagawa, T. (2018) Mining Educational Data to Predict Academic Dropouts: A Case Study in Blended Learning Course. *Proceedings of IEEE TENCON* 2018, Jeju, 28-31 October 2018, 2205-2208. https://doi.org/10.1109/TENCON.2018.8650138

[5] Jimenez, F., Paoletti, A. and Sciavicco, G. (2019) Predicting the Risk of Academic Dropout with Temporal Multi-Objective Optimization. *IEEE Transaction on Learning Technologies*, **12**, 225-236. https://doi.org/10.1109/TLT.2019.2911070

[6] Dharwaman, T., Ginardi, H. and Munif, A. (2018) Dropuot Detection Using Non-Academic Data. *Proceeding of* 4*th International Conference on Science and Technolgy* (*ICST*), Yogyakarta, 7-8 August 2018, 1-4. https://doi.org/10.1109/ICSTC.2018.8528619

[7] Burgos, C., Campanario, M.L., de la Peña, D., Lara, J.A., Lizcano, D. and Martínez, M.A. (2017) Data Mining for Modelling Students' Performance: A Tutoring Action Plan to Prevent Academic Dropout. *Computers and Electrical Enginnering*, **66**, 541-556. https://doi.org/10.1016/j.compeleceng.2017.03.005