Efficiency of Mining Algorithms in Academic Indicators

Efficiency of Mining Algorithms in Academic Indicators

Amelec Viloria[1], Hugo Hernández Palma[2], William Niebles Núñez[3], Mercedes Gaitán[4] , Bonerge Pineda Lezama[5]

[1]Universidad de la Costa, Barranquilla, Colombia
[2] Universidad del Atlántico, Puerto Colombia, Atlántico, Colombia.
[3]Universidad de Sucre, Sincelejo, Sucre, Colombia.
[4]Corporación Universitaria Empresarial de Salamanca (CUES), Barranquilla, Colombia.
[5]Universidad Tecnológica Centroamericana (UNITEC), San Pedro Sula, Honduras

[1]Email: jesussilvaUPC@gmail.com

Abstract. Data Mining is the process of analyzing data using automated methodologies to find hidden patterns [1]. Data mining processes aim at the use of the dataset generated by a process or business in order to obtain information that supports decision making at executive levels [2] [3] through the automation of the process of finding predictable information in large databases and answer to questions that traditionally required intense manual analysis [4]. Due to its definition, data mining is applicable to educational processes, and an example of that is the emergence of a research branch named Educational Data Mining, in which patterns and prediction search techniques are used to find information that contributes to improving educational quality [5]. This paper presents a performance study of data mining algorithms: Decision Tree and Logistic Regression, applied to data generated by the academic function at a higher education institution.

1. Introduction

In order to apply a data mining project, in any type of scenario, it is necessary to carry out a study of the available algorithms to determine the one that best suits the needs of the project to be carried out [6]. For this reason, a study has been developed on the performance of Microsoft's Decision Tree and Logistic Regression algorithms, applied to the academic data of a higher education institution.

Previous studies such as that of the University of Minho in Portugal [7], took data about high school students from two public institutions in the same country to apply prediction techniques. Three different mining purposes and four data mining methods were tested. The obtained results revealed that it is possible to achieve high precision in the prediction, given the data of two academic periods. At Awadh University in India, a student achievement study was conducted based on a group of 60 students from different careers [8]. The mining classification task was used on the basis of student data to predict student division. Information such as attendance, tests, seminars and assignments were collected to predict performance at the end of the academic period [9], [10], [11].

Data mining techniques come from artificial intelligence and statistics, and are embodied in algorithms which are then applied to a set of data to obtain results [12] [13]. Each algorithm is designed to accept or drop different types of data, so from that perspective, algorithms that do not accept the types of data existing in each project can be discarded [14]. In this scenario, due to the use of both discrete and continuous data, Microsoft's Decision Tree and Logistic Regression algorithms are chosen over the

others implemented by Microsoft's Data Tools Analysis Services (SSDT), a tool associated with the development of databases and business intelligence projects.

The Logistic Regression algorithm is a type of statistical analysis oriented to the prediction of a categorical variable in function of other variables considered as predictive parameters [15]. Specifically, the algorithm implemented by Microsoft turns out to be a variant of the neural network algorithm. This type of algorithm accepts any type of input and so it is considered flexible, and fits various analytical tasks within data mining, including prediction, classification, and exploring and weighing the factors that contribute to a specific outcome [16].

For their part, decision trees and rules that use invariant divisions have a simple form of representation, making the inference model relatively simple for the user's understanding [17]. A decision tree model has a unique primary node that represents the model and its metadata. Below the primary node are independent trees representing the prediction attributes selected.

A common variable to the algorithms described above is performance, because it is defined as the characteristic related to response time, resource usage (RAM and CPU) and reliability of operations [18]. The reliability of a prediction algorithm is given by the precision with which a resulting model defines the input data set. This factor is quantifiable with Microsoft's Data Tools analysis tool. The study seeks to determine the algorithm with better performance between: Decision Tree and Logistic Regression of Microsoft, on continuous and discrete data of academic indicators from a higher education institution.

2. Test Environment

The server on which the tests were performed presents the hardware and software features described in Table 1.

Table 1. Server specifications [19]

Characteristic	**Specification**
Processor	Intel® Core™ i7-4702MQ @2.20GHz
RAM	8GB
Hard Disk	1TB
Operating system	Windows 8.1
Database engine	Microsoft SQL Server 2012 Standard

This testing environment is oriented to cover the objectives related to the axes of the academic process which are: admission, enrolment, promotion and graduation. The requirements are listed below:

- Determine behavior patterns for student admissions; by career and faculty.
- Determine behavior patterns for student enrollment and subject selection (number of subjects, number of credits, level and subject area, etc.); by career, by level, and by faculty.
- Determine the factors that influence the cases of dropout (desertion and loss of the subject by attendance); by career, levels, subjects, areas of knowledge and faculty.
- Determine behavior patterns in the academic promotion of students; by subject, level, areas of knowledge, career, and faculty.
- Determine factors that influence second and third enrollment scenarios; by subject, level, areas of knowledge, career, and faculty.
- Determine the factors that influence the cases of students with low terminal efficiency; by career and faculty.

In order to measure indicators, the total number of mining models that meet the requirements was taken as the population. Because each requirement needs different levels of detail, 1025 models were considered, selected as the population number, and the sample size is defined by the formula [20].

Each of the algorithms was applied to a data structure that satisfies the defined requirements. These structures are a set of data obtained with SQL language and integrated to the Data Tools through the "Data View" option. The attributes used for each data structure are described in Table 2.

Table 2. Attributes of data structures

Indicators	**Attributes**
Entry	ID [string], names [string], age of registration [integer], city of provenance [string], province of provenance [string], country of provenance [string], sex [string], marital status [string], career of inscription [string], institute of provenance [string], title [string].
Enrollment	Key [integer], student code [string], names [string], nationality [string], sex [string], age of enrollment [integer], total subjects chosen [string], total credits chosen [string].
Dropout	Key [integer], subject [string], theoretical hours [integer], practical hours [integer], field of study [string], enrolment number [integer], attendance [integer], subject level [integer], student names [string], student sex [string], student nationality [string], student marital status [string], teacher names [string], teacher gender [string], teacher marital status [string], teacher nationality [string], teacher type [string], teacher title type, dropout form [string].
Promotion	Key [integer], subject [string], theoretical hours of the subject [integer], practical hours of the subject [integer], field of study [string], enrolment number [integer], subject level [integer], promotion note [whole], student names [string], student gender [string], student nationality [string], teacher names [string], teacher nationality [string], teacher gender [string], teacher marital status [string], teacher type [string].
Repeat	Key [integer], subject [string], theoretical subject hours [decimal], practical subject hours [decimal], subject area [string], enrolment number [integer], attendance [integer], subject level [integer], student names [string], student sex [string], student nationality [string], teacher names [string], teacher sex [string], teacher marital status [string], teacher nationality [string], teacher type [string], final note [decimal].
Terminal Efficiency	Key [integer], student names [string], student gender [string], nationality [string], graduation project area [string], grade point average [decimal], grade point average [decimal], final credits [decimal], age of entry [integer], time to finish pensum [integer], terminal efficiency [integer].

Scenarios arise when an algorithm is added to the data structure; that is, they are defined with the implementation of the logistic decision tree and logistic regression algorithm respectively. The steps to follow to carry out this process are [19]:

a) Select the definition of the origin of the data structure (relational).
b) Select the Logistic Regression algorithm for scenario 1 and Decision Tree for scenario 2.
c) Select the source data set.
d) Specify the types of data.
e) the learning and testing set for the algorithm.
f) Name the data mining model and structure.

3. Results and Discussions

3.1 Response Time

The response time was measured in seconds with the help of Microsoft's Data Tools. Figure 1 presents a summary graph of the results.

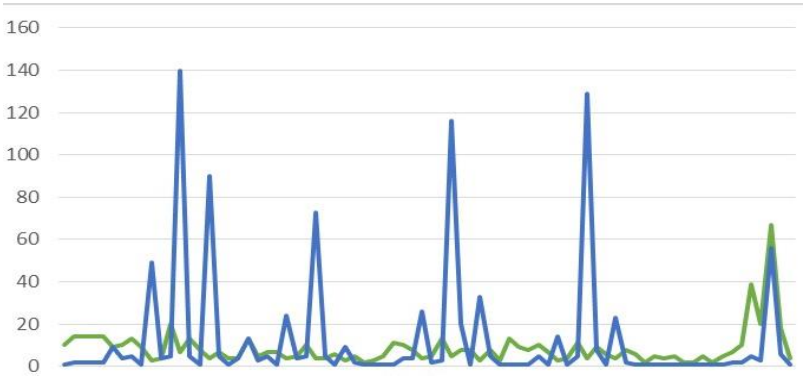

Figure 1. Response Time. Decision tree in blue and logistic regression in green.

The figure above shows that the response time values for the logistic regression algorithm have greater variation than the values for the decision tree algorithm. Descriptive data indicate that: the mean of the 86 sample values for the decision tree algorithm is 8.54 seconds, with a standard deviation of 8.58; versus 12.11 seconds for the logistic regression algorithm with a standard deviation of 26.14.

3.2 CPU usage

This indicator was taken in percent with the help of the Microsoft system monitor. Figure 2 presents a summary graph of the results obtained from this process.

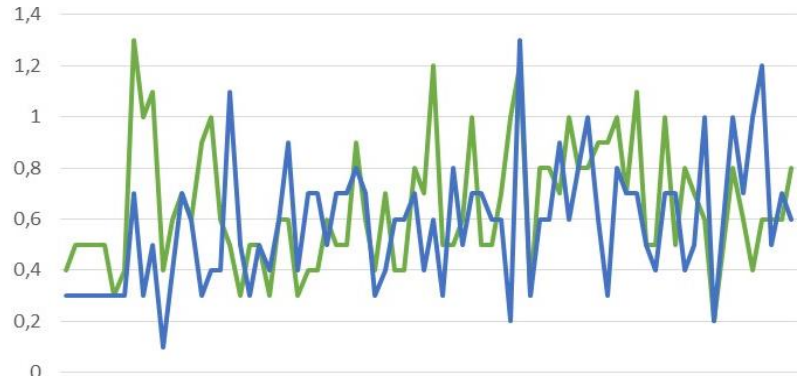

Figure 2. CPU Usage in Percent. Decision tree in blue and logistic regression in green.

The values obtained show great variability in the use of CPUs in the two study algorithms, with the mean of the 86 values taken for the decision tree algorithm being 0.61% with a standard deviation of 0.15; while the logistic regression algorithm has an average of 0.53% with a standard deviation of 0.20.

3.3 RAM Use

This indicator was taken in Megabytes with the help of the Microsoft system monitor. Figure 3 presents a summary graph of the results obtained from this process.

Figure 3. RAM usage in megabytes. Decision tree in blue and logistic regression in green.

The values with respect to the use of RAM shown in the previous graph do not show a significant difference between the two study algorithms, with the mean of the 86 values taken for the decision tree algorithm is 130.11 MB with a standard deviation of 21.22, while the mean for the logistic regression algorithm is 134.58 MB with 23.72 standard deviation.

3.4 Accuracy

The accuracy indicator is a value taken from Microsoft's Data Tools. Figure 4 presents a summary graph of the results obtained from this process.

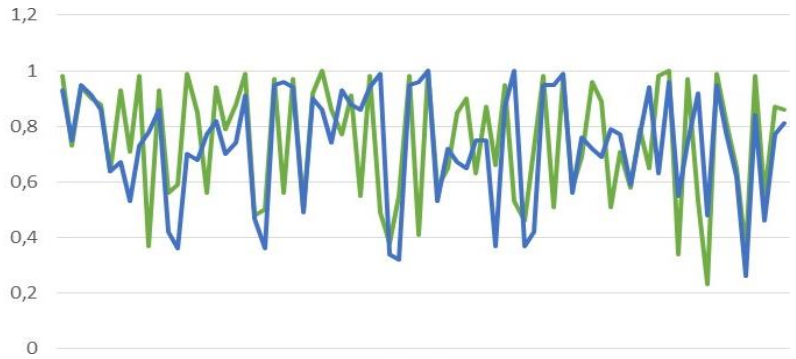

Figure 4. Accuracy in scoring. Decision tree in blue and logistic regression in green.

The accuracy values of the algorithms, shown in the graph above, show variability in the two case studies, which is confirmed by the descriptive data: the decision tree algorithm has an average of 0.79 in accuracy of the 86 models obtained with a standard deviation of 0.25; while the logistic regression algorithm has an average of 0.80 with a standard deviation of 0.21. The decision tree algorithm has an average of 0.79 in the accuracy of the 86 models obtained with a standard deviation of 0.25; while the logistic regression algorithm has an average of 0.80 with a standard deviation of 0.21.

3.5 Hypothesis Contrasts

In order to determine the best performance algorithm to be applied to the academic data of the higher education institution, four hypotheses were contrasted with respect to the defined performance indicators: response time, CPU use, RAM use and accuracy. The null hypothesis in the mean comparison tests for each indicator is that the performance of the two algorithms is the same; therefore, the algorithm with the highest performance can be determined in cases where the null hypothesis is rejected in these mean comparison tests.

Since the sample data set does not meet the normal hypothesis, the nonparametric test "Wilcoxon Sign Range Test" was applied for related samples. And the results are shown in Table 3:

Table 3. Hypothesis test summary

Null hypothesis (Ho)	Significance (4 tails)	Decision
The median of the differences between the Logistic Regression response time and the Logistic Tree response time Decision equals 0.	0,004	Reject null hypothesis
The median of the differences between the accuracy of Logistic Regression and the accuracy of Decision Tree is equal to 0.	0,052	Reject null hypothesis
The median of the differences between the use of Logistics Regression CPU and the use of Decision Tree CPU equals 0.	0,043	Reject null hypothesis
The median of the differences between the use of Logistic Regression RAM and use of the Decision Tree RAM equals 0.	0,047	Keep null hypothesis

From the comparison between sample means, for those variables in which Ho is rejected, it can be determined that there is sufficient statistical evidence to affirm that:

- The Decision Tree algorithm with 8.14 seconds exceeds Logistic Regression in response time with 12.10 seconds.
- The Logistic Regression algorithm has an average of 0.53%, which surpasses the Decision Tree algorithm, whose average is 0.61% with respect to CPU usage.
- The Decision Tree algorithm has an average accuracy of 0.71; while the Logistic Regression Algorithm has an average accuracy of 0.73, which defines that Decision Tree overpasses in the precision test.

Response time and accuracy results favor the Decision Tree algorithm, while Logistic Regression outperforms it in CPU usage. The use of RAM was similar for both algorithms. Since accuracy is the most important indicator for determining the best performance algorithm, the Decision Tree algorithm is chosen to be applied in the analysis of academic indicators in higher education.

4. Conclusions

The analysis of the performance of the Decision Tree and Logistic Regression algorithms, under the indicators of response time, CPU usage, RAM usage and accuracy, over academic indicator data, reveals that the accuracy of these algorithms is different. In this way it is established that the Decision Tree algorithm has better accuracy, due to the fact that its sample mean value is higher. It should be noted that the precision indicator is the most important for establishing the performance of a data mining algorithm.

The algorithms do not present significant difference in the use of RAM, which can be attributed to the fact that the algorithms were tested under the same data structures and the RAM storage required for the process is related to the amount of input data provided. The CPU usage of the Decision Tree and Logistic Regression algorithms is different and it is determined that Logistic Regression makes less use of this resource versus Decision Tree.

At response time, the Decision Tree algorithm has a lower mean than the Logistic Regression algorithm, so it can be concluded that it does it faster against the same scenario, because the second algorithm uses one analysis for each prediction attribute while the first one uses the input data as a single set for analysis.

Under the circumstances indicated in the previous paragraphs, the Decision Tree algorithm was determined as the algorithm with the best performance over academic data. Obtaining the socioeconomic data of the students will contribute with new patterns within the extraction of knowledge, so this research suggests to develop a plan to integrate this information to a study on mining algorithms.

References

[1] Han, Jiawei. Introduction to Data Mining. San Francisco: Morgan Kaufmann, 2006. págs. 1-20.
[2] Jain, Mugdha, and Chakradhar Verma. "Adapting k-means for Clustering in Big Data." International Journal of Computer Applications 101.1 (2014): 19-24.
[3] Huebner, Richard. A survey of educational data-mining research. Norwich: Norwich University, 2013. pág. 13.
[4] Maclennan, Jamie. Data Mining with Microsoft SQL Server 2008. Indianapolis, EEUU, Wiley Publishing Inc. 2008. págs. 39-53.
[5] Vallejos, Sofía. Minería de Datos. Corrientes, Argentina, Universidad Nacional de Noreste, 2006, págs. 11-16.
[6] Viloria, A. "Commercial strategies providers pharmaceutical chains for logistics cost reduction." Indian Journal of Science and Technology 8, no. 1 (2016).
[7] Viloria, A., & Gaitan-Angulo, M. (2016). Statistical Adjustment Module Advanced Optimizer Planner and SAP Generated the Case of a Food Production Company. Indian Journal Of Science And Technology, 9(47). doi:10.17485/ijst/2016/v9i47/107371.
[8] Karatzoglou A., Smola A., Hornik K. and Zeileis A. kernlab - An S4 Package for Kernel Methods in R. Journal of Statistical Software, 11(9), 1-20, 2004.
[9] N. Sapankevych y R. Sankar, "Time Series Prediction Using Support Vector Machines: A Survey", IEEE Computational Intelligence Magazine, vol. 4, núm. 2, pp. 24–38, may 2009.
[10] F. Villada, N. Muñoz, y E. García, Aplicación de las Redes Neuronales al Pronóstico de Precios en Mercado de Valores, Información tecnológica, vol. 23, núm. 4, pp. 11–20. 2012.
[11] Venugopal K, K.G. Srinivasa and L. M. Patnaik. Soft Computing for Data Mining Applications. Springer Berlin Heidelberg: Springer-Verlag. ISBN 978-3-642-00192-5, pp 354, 2009.
[12] F. Villada, N. Muñoz, y E. García, Aplicación de las Redes Neuronales al Pronóstico de Precios en Mercado de Valores, Información tecnológica, vol. 23, núm. 4, pp. 11–20. 2012.
[13] Chapman B, G. Jost and R Van der Pas. Using OpenMP: Portable Shared Memory Parallel Programming Scientific and Engineering Computation. The MIT Press.Massachusetts Institutte of Technology. ISBN 978-0- 262-53302-7. pp 349. 2008.
[14] Jain, Mugdha, and Chakradhar Verma. "Adapting k-means for Clustering in Big Data." International Journal of Computer Applications 101.1 (2014): 19-24.
[15] Ceruto T, O. Lapeira, A. Rosete and R. ESPÍN.Discovery of fuzzy predicates in database. Advances in Intelligent Systems Research (AISR Journal), vol. 51, No 1, pp. 45-54, ISSN 1951-6851, Atlantis Press, 2013.
[16] Amelec, V., & Alexander, P. (2015). Improvements in the automatic distribution process of finished product for pet food category in multinational company. Advanced Science Letters, 21(5), 1419-1421.
[16] Ruß G. Data Mining of Agricultural Yield Data: A Comparison of Regression Models, In: Perner P. (eds) Advances in Data Mining. Applications and Theoretical Aspects, ICDM 2009. Lecture Notes in Computer Science, vol 5633.
[17] Taylor S. and Letham B. prophet: Automatic Forecasting Procedure. R package version 0.1. 2017
[18] Wuo W., Xue H. An incorporative statistic and neural approach for crop yield modelling and forecasting, Neural Computing and Applications, 21(1): 109–117, 2012.
[19] Ji, B., Sun Y., Yang S. and Wan J. Artificial neural networks for rice yield prediction in mountainous regions, Journal of Agricultural Science, 145: 249-26, 2007.
[20] Karatzoglou A., Smola A., Hornik K. and Zeileis A. kernlab - An S4 Package for Kernel Methods in R. Journal of Statistical Software, 11(9), 1-20, 2004