IMCIC'18

The 9th International Multi-Conference on Complexity, Informatics and Cybernetics

March 13 - 16, 2018 – Orlando, Florida, USA

PROCEEDINGS

Volume I

Edited by:

Nagib Callaos
T. Grandon Gill
Jeremy Horne
Belkis Sánchez
Michael Savoie

IIIS

Organized by

International Institute of Informatics and Systemics

Member of the International Federation for Systems Research (IFSR)

COPYRIGHT

ISBN: 978-1-941763-68-1 (Collection) ISBN: 978-1-941763-69-8 (Volume I)

HONORARY PRESIDENT
William Lesso
(1931-2015)

PROGRAM COMMITTEE CHAIRS of IMCIC 2017
and/or ITS COLLOCATED EVENTS
T. Grandon Gill
Jeremy Horne
Michael Savoie
Hsing-Wei Chu
Sushil Archarya
Andrés Tremante

CO-CHAIRS OF PROGRAM COMMITTEES and/or ADVISORY BOARDS
Thierry Lefevre
Mohammad Siddique
Esther Zaretsky
Reza Sajjadi
Taha A. Al-Douri

GENERAL CHAIRS
Nagib Callaos
Hsing-Wei Chu

ORGANIZING COMMITTEE CHAIR
Belkis Sánchez

CONFERENCES PROGRAM MANAGER /
HARDCOPY PROCEEDINGS PRODUCTION CHAIR
María Sánchez

TECHNICAL CONSULTANT ON COMPUTING SYSTEM /
CD PROCEEDINGS PRODUCTION CHAIR

Juan Manuel Pineda

META REVIEWERS SUPPORT

Maria Sánchez

OPERATIONAL ASSISTANT

Marcela Briceño

CONTENTS

Volume I

ICT Applications in Education

ICT Applications in Health Care and Bio-Medical ICT

Educational Data Mining - An Experience in UTN FRRe

David L. LA RED MARTÍNEZ, Mirtha E. GIOVANNINI, María E. BÁEZ MOLINAS, Juliana I. TORRE, Nelson YACCUZZI
Resistencia Regional Faculty, National Technological University
Resistencia, (3500), Argentine

ABSTRACT

This paper proposes the use of Data Warehousing and Data Mining techniques on performance, social, economic, demographic and cultural data from students who took "Algorithms and Data Structures", which is a subject in the Information Systems Engineering curricula at UTN-FRRe (Resistencia, Chaco, Argentina), to establish generic academic performance profiles.

Keywords: Academic Performance; Educational Data Mining; Predictive Data Mining; Higher Education; Course Assessment; Student Assessment.

1. INTRODUCTION

Academic performance is linked to the correct assimilation of contents, other related activities and personal characteristics (social and individual), being a critical element of analysis since it may reflect many features of educational institutions. During the first years of university, the performance impacts strongly on the decision to continue or abandon the studies and that is why universities should focus their efforts on motivating and retaining students who, from the beginning, show shortcomings in their academic performance [1].

Academic performance can be measured by observing the marks obtained by students when their knowledge, skills and abilities are tested. However, hardly does this evaluation provide any information that can be used to detect and correct cognitive problems, apprehension, etc. This is the reason why other factors which affect, directly or indirectly, both social and financial situations should be considered, as well as prior educational experiences, and thus establish student performance profiles [2].

Considering that the performance during the first year serves as a very good indicator of the student's future academic career [1], this paper examines academic performance in the core subject of the first level of the Information Systems Engineering (ISE) career in UTN-FRRe, Algorithms and Data Structures (ADS). This subject has a very high rate of students that either failed or quit it, that is, students who will take the subject again because they failed at the several instances of evaluation. For this reason, there is a need to analyze the existence of socioeconomic and behavioral patterns that distinguish different profiles of academic performance.

One way to achieve this is to develop evaluation methods that take advantage of the capabilities of the information technologies available. In this way, Data Warehouse techniques (DW) and Data Mining techniques (DM) are extremely useful tools for obtaining knowledge in large volumes of data. A DW is a collection of data-oriented topics, integrated, non-volatile, variable in time, used to support the process of management level decision making [3]. A DM is the knowledge discovery stage in databases, which consists of the use of specific algorithms that generate a list of patterns from pre-processed data [4].

This paper proposes the use of Data Warehousing and Data Mining techniques on the performance, social, economic, demographic and cultural data obtained from students of Algorithms and Data Structure from Information Systems Engineering career in the Resistencia Regional Faculty, branch of the National Technological University (Resistencia, Chaco, Argentina), in order to establish profiles of academic performance characteristics. From the descriptive analysis obtained during the 2013-2015 school years in the class aforementioned, a predictive model is used. It establishes the possibility of students´ academic failure taking into account the factors mentioned above.

The article is structured as follows: in Section 2 DM basics are explained with emphasis on the techniques used. In Section 3, the design of the proposed model is displayed. In Section 4 the results obtained are shown; and finally, in Section 5, the conclusions are delivered.

2. DATA MINING

Data Mining is a process whose purpose is to discover, extract and store relevant information from large databases through search programs as well as to identify patterns and global relationships, trends, deviations and other indicators.

Many authors consider data mining an essential step in the process of knowledge discovery in database. It consists of an iterative sequence of steps [5]: data cleansing, to remove noise or irrelevant data; data integration, where multiple data sources can be combined; selecting data, where data relevant to the analysis task are retrieved from the database; transforming data, when the data are transformed or consolidated into appropriated forms for mining operations by, for example, performing summary or aggregation; data mining, an essential process in which intelligent methods are applied in order to extract data patterns; evaluation of patterns, to identify patterns of interest that represent knowledge based on some interesting measures; presentation of knowledge, where visualization techniques and knowledge representation are used to present the user´s knowledge extracted.

The discovered knowledge can be applied to decision making, process control, information management and query processing among others [6].

Data Mining Algorithms

Data mining algorithm is a set of calculations and heuristic rules that enables the creation of a data mining model from the data. To create a model, the algorithm first analyzes the data provided, looking for specific types of patterns or trends. The algorithm uses the results of this analysis to define the optimal parameters for creating data mining model. Then these parameters are applied over the data set to extract processable patterns and detailed statistics [7]. The following data mining algorithms are used for this work:

Decision Trees: this algorithm is one of the most supervised of the learning methods used. One of its main virtues is the simplicity of the models obtained [8]. The algorithm generates the model by creating a series of divisions in the tree. These divisions are represented as nodes. The algorithm adds a node to model every time an input column has a significantly correlation with the predictable column [7]. Basically, an attribute is chosen as the next level of the tree if it can help discriminate more objects and it tends, in fact, to reduce entropy [9]. A decision tree performs a test while it is traversed to the leaves in order to reach a decision. An internal node contains a test on a value of a property. A node indicates probability that a random event should occur according to the nature of the problem. This type of nodes is round whereas others are square. A leaf node represents the value returned by the decision tree. The branches provide the possible paths which are in agreement with the decision [10].

Demographic Clustering: provides a quick and natural grouping of large databases. The number of clusters to be created is automatically determined, and these distributions are characterized by the value of its members [11], [12]. The demographic clustering algorithm builds sets by comparing each record with all groups created previously by assigning the registration of grouping that maximizes a similarity score [13]. The quality of a partition is evaluated by a global measure, which favors groups with high similarity. In each step, the algorithm uses this criterion to decide whether to assign a record in an existing cluster, or to create a new one. This process ends when the results of the iteration are unchanged in the group [14].

3. PROPOSED MODEL

The performance profiles obtaining process consists of different phases: Data Collection, Data Warehouse Assembling and Data Mining (Fig. 1).

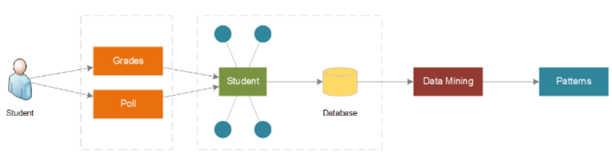

Fig. 1. Profile obtaining process.

Data Collection

The first phase involves the collection of the needed information that feeds the subsequent phases for determining the academic performance profiles. These profiles allow to relate students who have certain socioeconomic characteristics with a certain academic performance (success or failure).

First, to establish the profiles some information about previous academic and socioeconomic factors that may affect the student academic performance is required. These aspects were defined by the study group based on previous research. Some of the covered aspects are: high school, current residence, time spent studying, employment status of both parents and student, parent’s studies and considerations about the use of ICT.

Information regarding the student’s academic performance during the school year is also required, including the grades they obtained in the different exam instances and practical exercises. Here it was only considered the performance over the course and not the grades of final exams, since both instances are independent of each other, exceeding the scope of this work. Because of that, the participation of both students and professor is required in the obtaining phase. In order to obtain personal aspects, initially the students participated actively giving direct answers. To do so they completed an online quiz about their academic and socioeconomic status. The aspects covered were the factors mentioned above, which have, over the years, proven to influence academic development.

The students’ grades at the different instances of term exams and their final condition (promoted, regular or free) are provided at the end of the course by the Algorithms and Data Structures class.

DW Assembling

There might be incompleteness, inconsistencies and incoherencies throughout the data, because some values of many features do not have certain restrictions. Therefore, the data obtained in the previous step must be subjected to a process of purification: null fields removing or filling, typographical errors correcting, and integration with exam results. This process ensures consistency and coherence of the data loaded into the DW.

The structure of the DW model used is very simple, it consists only of a Student fact table and several tables for associated dimensions. The fact table includes student’s specific information, academic performance, and final situation, promoted, regular, or free in the course of analysis. The dimensions are the characteristics over study; they contain descriptive information obtained in the quiz about the student´s socio-economic background.

Data Mining

The phase of Data Mining comes after the DW assembling and loading. For this, some techniques were selected, creating related mining flows, which parameterize the respective algorithms.

The selection of the different algorithms was based on the advantages each one provided. Thence, the clustering technique (Demographic Cluster) was chosen. It allows to find useful characterizations for the building of classifiers, and it also enables the discovery of groups and subgroups to reveal the nature of the problem structure. The object of this technique is to obtain groups or sets from similar elements. The Decision Trees Classification technique was used in a greater extent.

Decision Trees are easy to use, they provide discrete and continuous attributes support, and process properly non-significant attributes and missing values. Their main advantage is the ease of interpretation, useful for high-dimensional problems: the problem is presented for analyzing all options. Their aim is to make classifications on known data and create models with them that can be used to predict or classify new or unknown values.

The previously described techniques were used to determine the performance profiles, they allowed a dimensional analysis of data considering the variable related to the final status of the student as mining parameter, determined by their status on the subject at the end of the school year (promoted, regular or free).

The results obtained were patterns that determine the data descriptive model, from which the performance profiles were estimated.

4. RESULTS

Dimensional Analysis of Students

In order to determine academic performance patterns, tests were made to the data of students over the 2013 to 2015 school years, with a total of 615 students. The main parameter taken for the analysis was their final condition at the end of the school year, that is, if the student was "regular", "free" or "promoted". It was considered "free" status of those students who did not pass the term exams or no longer attended the class. Students who "have regularized" the class, are those who passed the three term exams with a score greater than or equal to 60% but did not reach 75%. Finally, students who "have promoted the class" are those who have passed the three instances of exams with a score over 75%.

According to the data collected, from the 615 students analyzed, 60% are free, 28% regular and 12% are promoted.

Besides the final condition, the factors mentioned above - high school education, current residence, time spent studying, employment status of parents and student, parents’ education and consideration regarding the use of ICT are also taken into account to define the profiles.

Analyzing each of the socioeconomic and cultural factors and comparing them with the final condition of the student, the Decision Tree and Demographic Clustering algorithms were applied and the following result were obtained.

The greatest academic performance was reached by students from private religious high schools, with 14% promoted and 28% regular. This shows that students who come from private religious high schools have better academic performance compared to those coming from other educational institutions, which have the highest percentage of free students: 60% for public high schools, 61% for private entities, and 72% for others.

Considering the criteria weekly hours of study, the highest percentage of promoted and regular students is seen in groups who study over 10 hours per week, which reaches a total of 43% in each of them.

On the other hand, the highest percentage of students free appears in the group that study up to 10 hours per week, say 65%.

Considering the employment status of the students, and the number of weekly hours devoted to studying, it can be observed that the greatest academic success is among those who work up to 20 hours, 53%, and from 21 to 35 hours, 52%, while the largest percentage of free students is among those who are not working, 62%, and those working more than 36 hours, 64%.

In the criteria weekly working hours of the mother it is shown that the highest percentage of academic success (promoted and regular) corresponds to the group whose mothers work more than 36 hours per week, in total 48% (15% and 33% respectively). The highest rates of free students correspond to those whose mothers do not work or work from 21 to 35 hours per week, reaching 64%.

In weekly working hours of the father it is seen that the highest percentage of academic success (promoted and regular) is the group whose fathers work more than 36 hours per week, which reaches a total of 44%. The highest percentage of free students which has been obtained corresponds to those students whose fathers do not work, 64%, or work up to 20 hours, 68%.

Study importance criteria display that the highest percentage of academic success is the group that claim to give greater importance to study than to spending time with their families, which ascends to 58%. Compared with the highest percentages obtained from free students who stated that they give more importance to study than to having fun, which corresponds to 67%.

Considering the gender criteria, the highest percentage of academic success (promoted and regular students) corresponds to the female group getting 43% in total. The highest percentages of free students correspond to the male group, reaching 61%.

Concerning the residence of the student, the highest percentage of academic success (promoted or regular students) corresponds to those living independently, getting to 47%. Conversely, the highest percentages of free students correspond to those living with their families, reaching 63%.

Regarding the mother´s studies criteria, considering only those categories that include at least 40 students for a more representative analysis, the highest percentage of academic success (promoted and regular students) corresponds to the group whose mothers have completed higher studies, accounting for 50%. The highest percentages of free students correspond to the group whose mothers did not complete high school, 66%, or university, with 65%.

Concerning the father´s studies criteria, the highest percentage of academic success (promoted and regular students) corresponds to the group whose fathers completed their higher education - 46% - and university, corresponding to 43%. The highest percentages of free students correspond to the group whose fathers did not finish college, with 67%.

Taking into account the consideration of the ICT, the highest percentage of academic success is the group that considers that the ICT skills will be essential for professional practice, getting a 44%. The highest percentage of free students with a representative sample of students belongs to the group that believes that ICT are a present reality, reaching 66%.

Analyzing student's motivation to study, the highest percentage of promoted and regular students belongs to those seeking to learn thoroughly and succeed, with 13% and 31% respectively. The highest percentage of free students 70%, is seen in those who only seek to pass the course.

Finally, considering the medical insurance of the student, we can see that the best academic performance, with 49% of promoted and regular students, is situated among those who have their own medical insurance, while the highest percentage of free students, 62%, lies on those without a medical insurance.

Performance Profiles Obtained

The previous dimensional analysis provides the main features that mainly determine the profiles of academic performance, either success or failure.

Table 1 summarizes those patterns having been mostly observed according to the analysis dimensions.

5. CONCLUSIONS

This paper proposes a model which allows the definition of academic performance profiles using DW and DM techniques; these are based on "Algorithms and Data Structures" students' summary data during 2013, 2014 and 2015. The paper includes the students' final academic status and the socioeconomic, cultural and attitudinal influence of the environment to their studies, establishing possible profiles of academic success or failure.

The obtained profiles determine, generally, that those female students who study more than 20 hours per week, work up to 20 hours per week, whose parents work more than 36 hours per week and have complete superior studies, give more importance to study than spending time with family, reside independently, consider the ICT as an essential tool for professional practice, are seeking to fully learn and pass and have Medical Insurance have a tendency to succeed. On the other hand, those male students who study up to 10 hours per week, do not work or work more than 36 hours per week, whose parents do not work or work up to 20 hours per week and did not finish high school or college, give more importance to study than to having fun, still live with their families, consider ICT only as a reality, seek only to pass the subject and have no Medical insurance have a tendency to fail.

The determination of these profiles provides the ability to predict the students´ future academic performance from the knowledge of the factors that affect them, verifying the correspondence that each of these have with the previously determined profiles.

Table 1. Patterns of features for each performance profile.

Academic success	Academic Failure
Study more than 10 hours per week.	Study up to 10 hours per week.
Work up to 35 hours per week.	Do not work or work 36 or more hours per week.
Their mothers work 36 or more hours per week.	Their mothers do not work.
Their fathers work 36 or more hours per week.	Their fathers do not work or work up to 20 hours per week.
Give more importance to study than spending time with their family.	Give more importance to study than having fun.
They have medical insurance.	They have no medical insurance.
The highest percentage of academic success is for the female group.	The highest percentages of academic failure is for the male group.
Live independently.	Live with their families or in any unforeseen situation of residence.
Their mothers are college postgraduates or have higher non-university studies completed.	Their mothers are high school graduates or did not finish their college studies.
Their fathers are college graduates or have a higher non-university education.	Their fathers are high school graduates, did not graduate from university, or have no studies.
They consider the dominance of ICT is essential for professional practice.	They considered that ICT are a trending reality.
Their motivation to study is to learn and pass the course.	Their motivation is to study and pass the course.

Acknowledgement. This work has been supported by the Project: "Diseño de un Modelo Predictivo de Rendimiento Académico Mediante la Utilización de Minería de Datos", code UTI3808TC of National Technological University (Argentine). Special thanks to Dr. Marcelo Karanik for his important contributions to the paper, especially his review of the English version.

9. REFERENCES

[1] M. Oloriz, M. L. Lucchini, E. Ferrero, **Relación entre el Rendimiento Académico de Ingresantes en Carreras de Ingeniería y el Abandono de los Estudios Universitarios**. Mar del Plata, Argentina, 2007.

[2] V. Tinto, **Leaving College. Rethinking the Causes and Cures of Student Attrition**. Chicago: The University of Chicago Press, 1993.

[3] W. H. Inmon, **Data Warehouse Performance**. John Wiley & Sons. USA, 1992.

[4] U. M. Fayyad, G. Grinstein, A. Wierse, **Information Visualization in Data Mining and Knowledge Discovery**. Morgan Kaufmann. Harcourt Intl, 2001.

[5] J. Han, M. Kamber, J. Pei, **Data Mining: Concepts and Techniques**. 3rd edition. Morgan Kaufmann, 2011.

[6] J. R. Zaki, W. Jr. Meira, **Data mining and analysis: fundamental concepts and algorithms**. Cambridge University Press, 2014.

[7] Microsoft: **Algoritmo de árboles de decisión de Microsoft**. [Online] Available on:https://msdn.microsoft.com/es-es/library/ms175312(v=sql.120).aspx , 2017.

[8] A. Roche, **Árboles de decisión y Series de tiempo**. Tesis de Maestría en Ingeniería Matemática. Facultad de Ingeniería, UDELAR. Montevideo, Uruguay, 2009.

[9] M. Theophano, **Temporal Data Mining**. 1st Ed. Chapman & Hall/CRC, 2010.

[10] P. A. Vizcaino Garzon, **Aplicación de técnicas de inducción de árboles de decisión a problemas de clasificación mediante el uso de WEKA (Waikato Environment for Knowledge Analysis)**. Fundación Universitaria Konrad Lorenz. Facultad de Ingeniería de Sistemas. Bogotá, Colombia. 2008.

[11] IBM Knowledge Center: **Distribution-based Clustering**. [Online]. Available on:http://www.ibm.com/support/ knowledgecenter/ SSEPGG_9.5.0/com.ibm.im.model.doc/ c_distribution_based_clustering.html, 2017.

[12] X. Xu, M. Ester, H.-P. Kriegel, J. Sander, "A Distribution-Based Clustering Algorithm for Mining in Large Spatial Databases". **Proceedings of 14th International Conference on Data Engineering (ICDE'98).** Orlando, Florida, USA, 1998.

[13] R. Brause, E. Hanisch, "Medical Data Analysis". **First International Symposium. ISMDA.** Frankfurt, Germany, 2000.

[14] S. C. Manganaris, M. Zerkle, D. K. Hermiz, **A Data Mining Analysis of RTID Alarms**. IBM. 1999.