ISSN 1870-4069

Determining the Relationship of Admission Features and Finishing University Studies using Educational Data Mining and Information Visualization

Josué Figueroa González, Beatríz A. González Beltrán,
Silvia B. González Brambila, Lourdes Sánchez Guerrero

Universidad Autónoma Metropolitana, Azcapotzalco,
Mexico

{jfgo, bgonzalez, sgb, lsg}@azc.uam.mx

Abstract. The analysis of academic data looking for patterns of interest has increased in the last years. However, this is not a simple process, especially considering the amount of data that is generated in an educative environment. With thousands of data, it is complicated searching for patterns that can simplify or improve an analysis. However, humans have an innate capacity for understanding images better than data, using this ability simplifies the process of finding patterns for its later analysis. This work shows the analysis of admission features for determining their impact in finishing or not of university studies using Educational Data Mining and Information Visualization. Results show that first approaches obtained from visualizing the data and finding patterns simplify the process of determining the effect of these features. It is also shown that a student with an older admission age is prone for not finishing their studies.

Keywords: admission features, educational data mining, information visualization, student performance, visual analytics.

1 Introduction

Educational data mining (EDM) refers to the use of Data mining techniques for analyzing academic or educational information [6]. Applications of EDM cover from predicting the marks in a single exam, to the creation of models for determining if a student would finish or not its studies. A traditional process of EDM involves the same steps of Data Mining (DM): obtaining data, a process for cleaning it, applying DM algorithms according to a desired goal and exploiting results for making decisions.

In DM process, the most important part are the data, they are the source of all the process, so knowing the data, their characteristics or patterns becomes a fundamental step at the moment of the analysis. As data grow, they become

pp. 27–35; rec. 2019-09-01; acc. 2019-09-24

more complex for analyzing, so having a previous comprehension of them before the analysis is desirable.

One of the challenges is how to analyze a set of data composed of thousands of registers with several characteristics for finding patterns that later can be studied or verified using EDM techniques. Information Visualization offers an answer to this; humans have a cognitive perception that makes easier finding information in a visual way rather than in a textual one. Information Visualization can be defined as the techniques used to represent textual information in a graphical way for facilitating its analysis [7].

Combining DM and Information Visualization creates a new area called Visual Analytics (VA) which is defined as the science related to analyzing information using visual interfaces [2]. In VA, users view graphic representations of the data and adjust parameters of models according to the interpretation of the graphics. Models are updated, and users can interact again with them. Information Visualization and DM combination applied to an educational environment is the focus of Visual Learning Analytics (VLA) [9].

Many public universities in Mexico face the problem that their rate of graduated students is not as high as expected, in some universities is less than 10%. Reasons for not finishing undergraduate level studies are several, can be personal, academic or labor ones. In the academic, can be identified: performance over different courses, but also, could be related to the admission characteristics of students.

This work presents a combination of Information Visualization and Educational Data Mining for determining the influence of the admission characteristics of entrance age, high school average and mark in the admission exam over students' completion of their professional studies. Goal of the work is showing the benefits of combining Visualization Information with EDM for simplifying the process of analyzing data and identifying those factors that affect students in finishing or not their studies.

2 Related Work

Analysis of whether a student will or not finishing its studies use mainly predictive algorithms and consider performance over different scholar periods in students' trajectory, personal characteristics of the students and admission features.

In [4] was performed a study of 8 different data mining algorithms to predict the performance of students who completed a module in a computing degree course. The sample was composed by the information of 22 students, which was obtained from the Student Information System (SIS) and the time spent by students on the Moodle platform. As a result of this analysis, it was determined that the Random Forest algorithm was the most suitable algorithm to predict the performance of students using the lower mean absolute error and the relative absolute error as evaluating metrics.

In [8], was proposed a Triadic Model for Teaching Analytics composed by a Teaching Expert, a Visual Analytics´ Expert and a Design-Based Research Expert that analyzed, interpreted and acted in real-time and in-situ, generating student´s learning activities to improve the learning environment. They promoted that to deal with the demands of the "New Demands on Teachers in the 21th Century Classrooms" project, teachers need to react in real-time and in situ for capturing information about student's learning, interpreting it, following the curricular goals and making reasoned decisions about next learning steps.

In [1], was studied the relationship between the cognitive admission entry requirements and the academic performance of students in their first year. The data set analyzed was composed of 1,445 student records from 2005 to 2009. Examined features were: students' entry age, the Joint Admissions and Matriculation Board score, the university score and the aggregate West African Examination Council score to predict the class of the student's first year grade. Were used six data mining algorithms. Logistic Regression had the highest prediction accuracy with 50.23% and Decision Tree had the least accuracy with 39.631%. The data mining models and regression shown that although there is a relationship between the admission requirements and the academic performance of students in their first year student, this is not very strong.

In [3], was performed a comparative analysis of data mining tools used for predicting student performance, highlighting the advantages of classification algorithms for this prediction.

Student's GPA is the main attribute that determine the performance of the student.

3 Data and Analysis Process

Analysis process involved discovering patterns through visualization techniques and then verify this patters using predictive algorithms. Were considered the general steps for applying data mining and visualizing information: data processing, visualizing, establishing a statement and finally, verifying the statement.

3.1 Data Set and Data Cleaning

Data set considered for this work was obtained through the General File of Students (AGA, for its Spanish acronym) which contains a lot of information about all the students that have or are studying at university. Were considered engineering students that entered university from the years 2010 to 2014, this period was considered for having a balanced set of students that had and had not finished their studies. The amount of students processed was 8,638. The considered admission characteristics were:

- Entrance age (AGE). Age of the student at the moment of entering university, with a range of years old from 16 to 58 with an average of 26.8.
- High school level average (AVG). The average of high school level obtained from the student, values run from 7 to 10 with an average of 8.2.

- Mark in the admission exam (EXA). The total of points obtained by a student in the admission exam, values run from 400 to 1000 for a maximum of 1000, the average of this feature was 677.

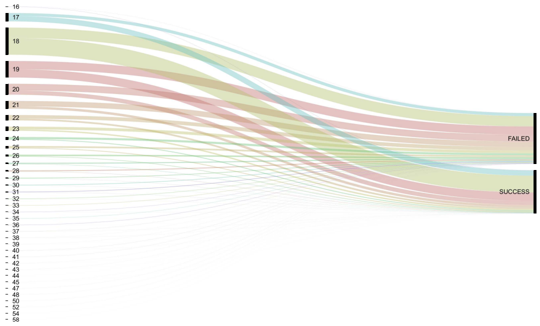

Fig. 1. Relationship of entrance age (AGE) and finishing or not university studies.

Output had two categories, if the student finished (Success) or not (Failed) their studies. Success was considered if the student had obtained its degree or had completed the total of credits of its college career. Failed status was given to those students that did not finish their studies and were dismissed because of reaching the limit time for finishing their studies (12 years) or due to failing the same course more than 5 times.

As part of the cleaning process, some criteria were transformed; for avoiding a lot of variations in the values, was used only the integer part of the average of medium high school level (AVG). Something similar happened with the mark in the admission exam (EXA), where students were grouped in hundreds, from 4 to 5 hundreds, 5 to 6, and so on until the limit of 1,000.

3.2 Visualizing and Analyzing Data

For presenting the relationship between each criterion and the result of Success and Failed, was used the Alluvial diagram using RawGraphics web platform [5]. In this graphic, there are weighted flows over a series of steps composed of nodes, and the size of flows is given by the amount of registers with the same characteristics.

All figures present in the left side the possible values of the displayed features, and in the right side is the result of finishing (Success) or not (Failed) university studies.

The relationship between entrance age and finishing or not university is shown in Fig. 1. Ages run from 16 to 58 years old. There is a clear relationship between the entrance age and the success or failure in the studies. Graphic shows that older students tend to fail in finishing them.

Fig. 2 shows the relationship between the average in the high school and finishing university. Integer average run from 7 (the minimum requested in the university) to 10, values of 0 and 6 appears due to some special cases. Students with an average in the range of 7, tend to fail in their studies. 8 can be considered as a break point with a similar distribution for each case. Students with an average of 9 and 10 tend to finish their studies.

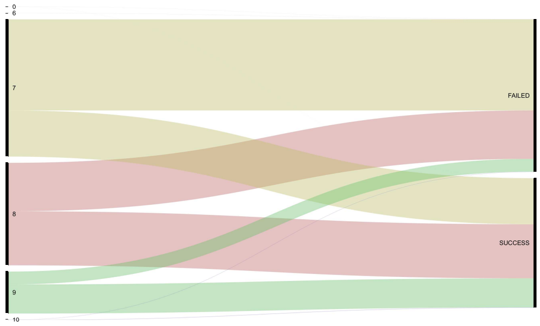

Fig. 2. Relationship of medium high school level average (AVG) and finishing or not university studies.

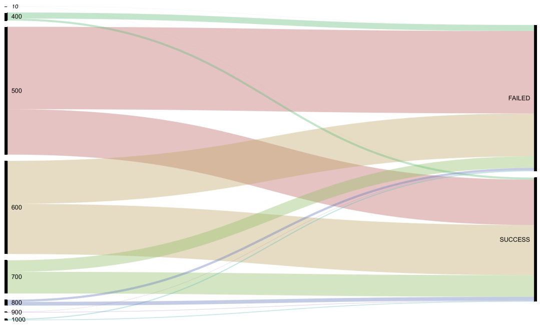

Fig. 3. Relationship of mark in admission test (EXA) and finishing or not university studies.

Fig. 3 presents the relationship of the mark in the admission exam and finishing or not the studies. Admission exam values run from 400 to 1,000; however university ask for a minimum of 500, but in some scholar periods, according to students requests, they can be accepted with a lower mark. Here can be noticed that students in the range from 400 to 500 points tend to fail in their studies. Again, there is a breaking point, 600 points, where there is a similar distribution for finishing or not. Students from 700 to 800 tend to finish their studies in a bigger proportion than those who do not. Students with the best marks, 900 and 1,000 are balanced in the amount of those that finish or not.

3.3 Establishing an Statement

Visual information presented in Fig. 1 to Fig. 3, shows that all criteria has a relationship with their value and finishing or not the studies. However, the one that has the most significant influence in finishing or not studies is entrance age. For this reason, were analyzed students grouped by ages.

- 16 to 18 years old students, where the proportion of students that finished their studies are greater than those who did not.
- 19 years old students appears to be the "breaking point", a proportion of 50% in each case it seems to be present.
- Students older than 20 years tend to have a bigger proportion of those that did not finish their studies. Although the number of students decrease, the trend of failing in their studies raises.
- This group was divided in other age ranges, considering those who have a similar performance, so were created: students from 20 to 23 years old, from 24 to 27 years old, and finally, students older than 28 years old.

From the data observed in the graphics, it was proposed the following statement: "entrance age has the greatest impact for a student for finishing or not their studies". This statement was verified using DM techniques. In this work, it was employed a predictive model considering that as bigger the accuracy, more true has to be the statement.

3.4 Verifying the Proposed Statement

For verifying the previous statement, it was generated a predictive model using 70% of the data for training. A classification model, considering all ages, was obtained using a decision tree, specifically the Classification and regression tree algorithm (CART). The model was tested then in different groups of ages.

- Considering only the AGE criteria.
- Considering AGE and AVG criteria.
- Considering AGE, AVG and EXA criteria.
- Not considering AGE.

According to the proposed statement and the figures, it was expected that a set with an older range of ages, has a bigger accuracy on the number of students classified as failed in their studies.

4 Results

The generated tree is shown in Fig. 4. As it can be seen, the root, the most important feature in a decision tree, is the entrance age (AGE) and contains the value of 20 years old. According to Fig. 1, the breaking point was 19 years old, meaning that 20 years old students tend to fail in their studies.

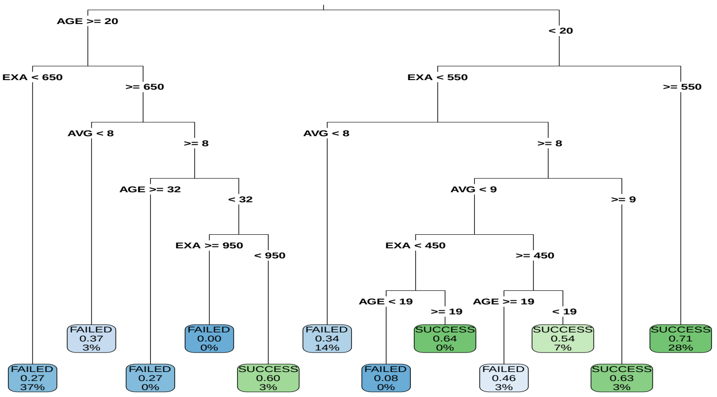

Fig. 4. Decision tree generated after processing data.

Table 1 contains the average accuracy for each group of age and different combinations of criteria.

Table 1. Accuracy of groups of ages and features.

Age group	(AGE)	(AGE+AVG)	(AGE+AVG+EXA)	Not AGE
All ages	60.4	59.7	58.3	58.3
16 to 18	62.51	63.91	67.49	65.84
19	50	60.52	64.60	63.5
20 to 23	66.04	66.04	67.55	65.5
24 to 27	73.02	73.02	74.34	69.7
28 or older	84.87	84.87	84.87	71

Table 1 shows that as the entrance age increases, the accuracy of the models raises, except in the case of 19 years old, the one that was identified as a break point in Fig. 1. This is because students younger than 20 years old did not show a clear distribution for finishing or not their studies, as shown in Fig. 1, meanwhile students older than 20 years old shown a clearer distribution.

5 Conclusions and Future Work

This paper presents the combination of Information Visualization and Educational Data Mining techniques for analyzing the effect of some admission characteristics over the undergraduate student successful rate. According to the state of art, a combination of Information visualization and Educational data mining can simplify the analysis process through finding patterns in a simpler way using images, then testing these patterns through data mining techniques.

Results show that the patterns found in the visualization stage were validated using a decision tree. The proposed statement was that students with an older age tend to not finishing their studies, and graphics show that score in admission test and average in high school have some importance. Visualization allowed determining the set of ages for which the generated model was more effective. Also, a predictive model from which results considering all ages was not efficient (accuracy of 60%) shows a good performance (more than 84%) for a group of older ages. For ages younger than 23 years old, the accuracy is very low. This means that for these students more features should be considered for predicting if they will or not finishing their studies. It was established that the entrance age to be affected was over 20 years old, the model shown this, but results shown that accuracy becomes significant at 24 years old.

This means that probably the group of 20 to 23 years old should be regrouped again. As future works, it is considered a more specific analysis on these groups of ages. Also a similar analysis using visualization should be applied to groups which have a low accuracy adding other characteristics more than admission ones.

References

1. Adekitan, A.I., Noma-Osaghae, E.: Data mining approach to predicting the performance of first year student in a university using the admission requirements. Education and Information Technologies pp. 1–17 (2018)
2. Cook, K.A., Thomas, J.J.: Illuminating the path: The research and development agenda for visual analytics. Tech. rep., Pacific Northwest National Lab.(PNNL), Richland, WA (United States) (2005)
3. Dagim, S., Proo, P., Pro, A.: Predicting performance and potential difficulties of university student using classification: Survey paper. International Journal of Pure and Applied Mathematics 118, 1314–3395 (2018)
4. Hasan, R., Palaniappan, S., Raziff, A.R.A., Mahmood, S., Sarker, K.U.: Student academic performance prediction by using decision tree algorithm. In: 2018 4th International Conference on Computer and Information Sciences (ICCOINS). pp. 1–5. IEEE (2018)
5. Mauri, M., Elli, T., Caviglia, G., Uboldi, G., Azzi, M.: Rawgraphs: a visualisation platform to create open outputs. In: Proceedings of the 12th Biannual Conference on Italian SIGCHI Chapter. p. 28. ACM (2017)
6. Romero, C., Ventura, S.: Educational data mining: a review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 40(6), 601–618 (2010)

7. Shneiderman, B.: The eyes have it: A task by data type taxonomy for information visualizations. In: The craft of information visualization, pp. 364–371. Elsevier (2003)
8. Vatrapu, R., Teplovs, C., Fujita, N., Bull, S.: Towards visual analytics for teachers' dynamic diagnostic pedagogical decision-making. In: Proceedings of the 1st International Conference on Learning Analytics and Knowledge. pp. 93–98. ACM (2011)
9. Vieira, C., Parsons, P., Byrd, V.: Visual learning analytics of educational data: A systematic literature review and research agenda. Computers & Education 122, 119–135 (2018)