Prediction of academic performance using data mining in first year students of peruvian university

Predicción del rendimiento académico mediante minería de datos en estudiantes del primer ciclo en una universidad peruana

Recibido: junio 22 de 2018 | Revisado: julio 26 de 2018 | Aceptado: agosto 02 de 2018

Eiriku Yamao[1]
Luis Celi Saavedra[2]
Rosalvina Campos Pérez[3]
Valery de Jesús Huancas Hurtado[2]

Abstract

Academic performance is a subject that has been studied for a long time. First year students in universities are the most vulnerable to face performance problems, resulting in possible desertion. Data mining in education applies data mining techniques in the information generated in the education sector. The present research consists of making the prediction of the academic performance of the students who entered the Professional School of Computer and Systems Engineering of the University of San Martín de Porres in the first cycle using data mining. Data were extracted from 1304 entrants who were classified using three factors: social, economic and academic, and predictions were made using three techniques: linear regression, decision tree and support vector machines, having the best result of 82.87% obtained using the decision tree. Out of the different factors, those that most influenced the academic performance were the following: admission exam grade, gender, age, income and distance from home to the study center. Using data mining it was possible to elaborate predictions of the academic performance of the students, which allowed the detection of students who could encounter issues in their studies during the first semester.

Key words: Academic Performance, prediction, Educational Data Mining, EDM, Higher Education

Resumen

El rendimiento académico es un tema estudiado desde hace mucho tiempo. Los alumnos ingresantes de las universidades son los más vulnerables a enfrentar problemas de rendimiento, resultando en posible deserción. La minería de datos en educación aplica técnicas de minería de datos en la información generada en el sector educación. El presente trabajo consiste en realizar la predicción del rendimiento académico de los alumnos que ingresaron a la Escuela Profesional de Ingeniería de Computación y Sistemas de la Universidad de San Martín de Porres en el primer ciclo utilizando minería de datos. Se extrajeron datos de 1304 ingresantes que fueron

1 Universidad de San Martín de Porres. Lima - Perú
eyamao@usmp.pe
2 Universidad de San Martín de Porres. Lima - Perú
lcelis@usmp.pe
3 Universidad Nacional Federico Villarreal. Lima - Perú
rcampos@unfv.edu.pe
4 Universidad de San Martín de Porres. Lima - Perú
vhuancash@usmp.pe

https://doi.org/10.24265/campus.2018.v23n26.05

clasificados en tres factores: sociales, económicos y académicos y se realizaron predicciones a través de tres técnicas: regresión lineal, árbol de decisiones y support vector machines, y el mejor resultado de 82.87% se obtuvo utilizando árbol de decisiones. De los diferentes factores, los que más influyeron en el rendimiento académico fueron los siguientes: nota de examen de admisión, género, edad, modalidad de ingreso y distancia desde su casa hasta el centro de estudios. Utilizando minería de datos fue posible realizar predicciones del rendimiento académico de los ingresantes. Esto permitió la detección de ingresantes que podrían enfrentarse a problemas en sus estudios.

Palabras clave: rendimiento académico, predicción, educational data mining, educación superior

Introduction

One of the main challenges in education is the exponential growth of data generated by information systems and technology and its use to improve the quality of educational services offered with a better decision making. Educational Data Mining is an emerging discipline that aims to take advantage of the new capabilities of data processing and the maturity of data mining algorithms to enhance the learning process and transform existing information into knowledge. (Han, Kamber & Pei, 2012) (Romero & Ventura, 2012) (Chalaris, Gritzalis, Maragoudakis, Sgouropoulou, y Tsolakidis, 2014) (Romero, Ventura, Pechenizkiy, y Baker, 2011).

The main subject of study in higher education is the academic performance of students. It´s not only a tangible value to measure the progress of a student in a given course or subject but it´s also one of the feature to see the level of success during and after obtaining a degree. Is of great importance for the educational institutions given that the level of success of their students is a reflection of the quality of the institution. (Calisir, Basak, y Comertoglu, 2016) (Rodríguez y Arenas2016) (López Bonilla, López Bonilla, Serra, y Ribeiro, 2015) (York 2015) (Shahiri y Husain, 2015).

Prediction is one of the oldest applications of EDM. Multiple studies have successfully created models to predict academic performance. It is a complex process given that multiple elements have been attributed to impact academic performance of students. A proper prediction model can be used to detect those who might face difficulties in their studies and be in risk of dropping out. Measures to help those students in risk by additional tutoring, changes and improvements in courses or curricula are some of the adjustments made previously. (Ramesh, parkavi y ramar, 2013) (Mishra, Kumar y Gupta, 2014) (Strecht, Cruz, Soares, Merdes-Moreria y Abren 2015) (ElGamal, 2013).

The most vulnerable students in universities, who might face difficulties and drop out are found in first year students. Adaptation to life in a university can be a great challenge for many and

some never manage to adapt completely. Multiple studies have found that the number of students with failing grades are higher in the first year, which extends their time in the university and in some cases dropping out. Student retention and their graduation are an important goal for universities, especially in STEM majors where student dropout rate might exceed the 30% in first year students. (Baradwaj & Pal, 2012) (Sepehrian, 2012) (Li, Rusk y Song, 2013) (Cheewaprakobkit, 2013).

Previous studies of prediction using EDM include Mishra et al. (2014) to predict academic performance of third year student in computer science major using J48 and Random Tree algorithm. Performance in previous semesters and other courses, leadership and motivation are influential in academic performance. Elakia & Aarthi (2014) uses student characteristics in high school to predict in which major they will have the best performance and discipline in universities based on behavioral trends in high school. Pal (2012), predicts low achievement and chances of dropout using multiple decision tree algorithm. Ramesh et al. (2013) uses prediction in final exam grade to find those who could fail a course, being occupation of the parents a strong impact in the results. Gray, McGuinness y Owende (2014) predicted academic achievement of first year student, using gender, age, high school grades, personality, motivation and learning style. Students under 21 years had a better prediction results. Sembiring, Zarlis, Hartama, Ramliana, y Wani (2011) uses SVM ad clustering to predict academic performance using interest, beliefs, family support, attitude and time spent studying achieving a high prediction rate of 93.67%.

There is very little studies made in Peru related to academic performance and almost none using EDM. This study is one of the first to create a prediction model from student characteristics found in Peruvian universities.

Materials and method

Methodology

The present study used a quantitative method. Scientific studies and related papers where searched to determine the proper variables needed to achieve the objectives of this study and to create a theoretical relationship between the variables. The design of this study is correlational – causal, aimed to describe the relationships between academic achievement and characteristics of first year students used to validate the prediction methods.

Population and sample

The population of this study are the first year students of the Information Systems career of the San Martin de Porres University. The sample has been taken from the years 2010 to 2015, giving 1304 students who were admitted, from which transfer students (who did not take the courses of the first year), those who dropped out in the first weeks or never enrolled in any course were removed from the data.

Dataset

All available data from admissions office and faculties has been collected and cleaned to create the dataset. From the revision of the literature and the

data found in the data repositories of the university, the selected variables for the study are:

- AGE. It´s the age of the student at the date of the first class in university.
- GENDER. The gender, male or female of the student, for this study a binary representation, zero (0) for female and one (1) for male is used.
- PROVINCE. To identify those students that came from another province to study in the capital.
- SCHOOLTYPE. Different types of school (national, private, religious and others).
- COLEEXC. Based on a list of the top 500 schools with student having the best grades in university.
- ADMISSIONEXAM. Grade of the student in the admission exam. Presented in percentage of the total score.
- DISTANCE. The total distance calculated from the place of residence of the student to the faculty where the classes were given. In some cases, there was no data for the address so the average distance from the selected district is used.
- APPROVED. To indicate the PASS or FAIL status of the student in the first year of university.

Data Mining Methods

Three data mining methods has been selected to be applied in the dataset.

Regression: The purpose for this model is to fit the data to a model based on variables. Answers questions like: what is the forecast of sales for the next month? (Han et al. 2012)

Decision Trees: Represent a group of classification rules in shape of a tree, based on an if-then ruleset. (Han et al. 2012).

For this study C5.0 algorithm was selected.

Support Vector Machines: Support vector machines is a type of algorithm that builds a model to represent simple point in a higher dimension to define a hyperplane to be used to create an optimal separation between classes to achieve proper classification. For the definition of the hyperplane, the algorithm uses the support vectors to map the data in a high enough dimension to make the classification. (Lantz, 2013)

Ethical aspects

The main ethical aspects about this type of research is the privacy of the personal data about students and professors. Information that can be used to identify a person requires authorization before its use and publication.

To assure the privacy of the data used in this study an anonymization process was applied to the data related to students and professors. A unique ID was assigned to each row of the data and every column that could be used for individual identification like first name, last name, ID card number, address has been removed. This way, protecting the privacy and the validity of the study results is achieved.

Result and discussion

Results

Regression

Linear Regression models aim to fit into a linear model y=f(x) all the attributes of the database don the relation between the dependent variable (y) and the independent variable (x). For this study, a logistic regression model has been used to create a model to predict the PASS/ FAIL condition of the first year students.

For the selection of the most influential variables to be added to the model, backward selection method has been used. The variables with the best fit are shown in Table 1. As expected admission exam score is the most important variable and the distance has a negative coefficient, meaning that student with less travel distance from their homes are more likely to pass the courses. The gender also has a negative coefficient, indicating that female student are more likely to earn a passing grade.

Table 1
Results from logistic regression

Coefficients:					
	Estimate	**Std. Error**	**Z value**	**Pr(>|z|)**	**Significance**
(Intercept)	-3.37717	0.87575	-3.856	0.000115	***
ADMISSIONEXAM	2.70438	0.47570	5.685	1.31e-08	***
AGE	0.10228	0.04379	2.335	0.019525	*
GENDER	-0.55567	0.24382	-2.279	0.022663	*
DISTANCE	-0.03678	0.01665	-2.209	0.027194	*

To use this model for prediction, the dataset has been split randomly into 75% for the training and 25% for the testing. The results of the prediction used this model is as seen on Table 2. Further examination of the results from this model has shown that there is a significant difference in the rate of passing the courses on the type of admission exam taken.

Table 2
Prediction results logistic regression

Exactitude	Sensibility	Specificity	AUC
67.4%	69.44%	59.45%	68.82%

Considering this new discovery, the type of admission exam was used to perform a further analysis and a new model with a better prediction capability was discovered using the variables in Table 3. In this new model, only considering those who were admitted via the ordinary type of admission process is considered. The results from the prediction using this model is as seen on Table 4.

Table 3
Results from logistic regression filtered by ordinary admission type

Coeficientes:					
	Estimate	**Std. Error**	**Z value**	**Pr(>|z|)**	**Significance**
(Intercept)	-4.2345	0.5179	-8.176	2.96e-16	***
ADMISSIONEXAM	10.9480	1.3225	8.278	< 2e-16	***
GENDER	-0.8137	0.3398	-2.395	0.0166	*
COLEEXC	-1.1600	0.5839	-1.987	0.0470	*

Table 4
Prediction results logistic regression filtered by ordinary admission type

Exactitude	**Sensibility**	**Specificity**	**AUC**
74.4%	74%	76%	82.12%

Decision trees

The C5.0 algorithm was used to create models based on an IF THEN rules to study the importance of each variable and predict the PASS / FAIL outcome. Boosting techniques has been applied to make multiple iterations of the algorithm to enhance the results. Using what was learned from the regression model, the type of admission was used a criteria to study the data. The best model found using the decision trees method is as seen on Figure 1.

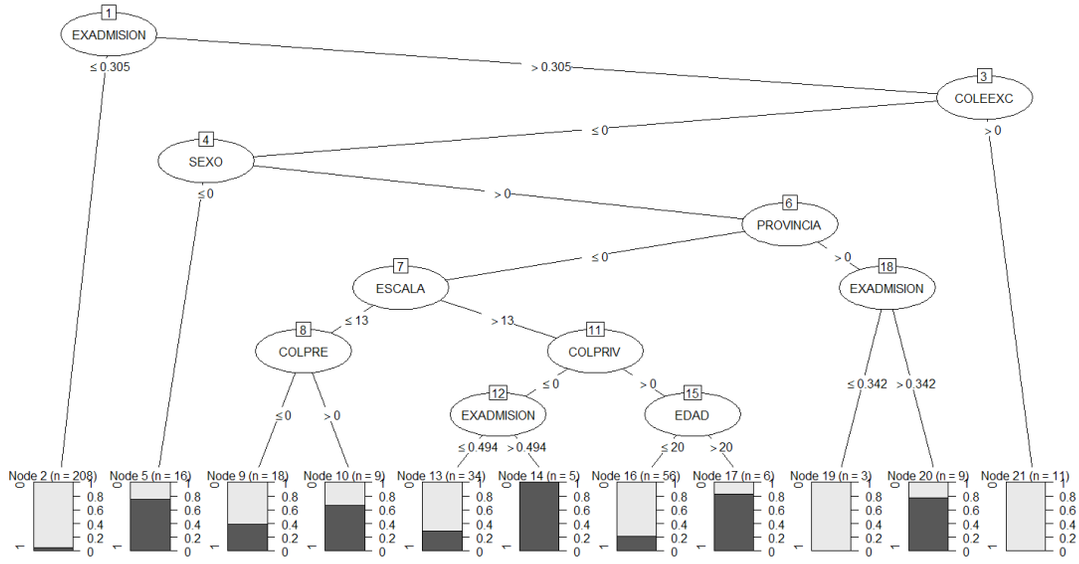

Figure 1. Decision Tree model filtered by ordinary admission type

Admission exam score, gender, age and variables related to the type of school are considered in this model. As in the regression model, a high admission exam score and being female are strong indicators of having a PASS outcome. Prediction using this model achieved an exactitude in the prediction of 82.87%.

Support Vector Machines SVM

Support vector machines method was used to create models based support vectors, to create a hyperplane to make the classification. An important part of SVM is the selection of the kernel to be used to model the data. The linear and polynomial kernel couldn't create a model good enough to be used for prediction, but the Gaussian and sigmoid kernels managed to create models, achieving the best results in predicting the outcome the Gaussian kernel with a exactitude of 75.2%

Discussion

Prediction of the outcome of the first year students has been achieved. The prediction results of the best models for each DM method used in this study is as seen on Table 5.

Table 5
Results of exactitude of prediction from Logistic regression, C5.0 decision trees and SVM algorithms

Algorithm	Prediction
Logistic Regression	74.4%
C5.0	82.87%
SVM	75.2%

As in other studies (Ecklund, 2013) (Li, Swaminathan, & Tang, 2009) (Veenstra, Dey, & Herrin, 2008) (Honken & Ralstron, 2013) (Elakia & Aarthi, 2014) (Gray et al. 2014), admission exam score is an important evaluation tool for prediction. The higher the admission exam score, the more likely for the first year student to pass the first semester.

Gender was also found to be important, as female students had a higher chance of passing the first semester and the closer the distance between the student's place of residence and the university, the more likely it is to have a passing grade.

The type of elementary and junior high school attended before studying in the university had little impact on student outcome.

Conclusions

Predicting the future outcome of first year students is an important way for universities to detect those who will most likely face problems to pass the first semester of studies, to give them the necessary support and prevent student dropout.

Female students in this study were found to have a higher passing rate than their female counterpart. Usually, being a gender minority, as is usual in STEM careers is considered to have a negative impact (Bayer, Bydzovska, Geryk, Obsıvac y Popelinsky, 2012) (Ecklund, 2013), which is a topic for future studies.

The type of school in elementary and junior high school did not have a strong

impact in the outcome, and the students coming from schools that are considered to the better ones either. It is most likely that other characteristics like emotional maturity, motivation, group of friends and others have a similar or stronger impact than academic ones.

As it is often mentioned, the data quality and quantity is important in these type of studies. Given that this study took place in one career in one university, it is of interest for future studies to compare the results with other careers like accounting (Byrne & Flood, 2008) or engineering (Li et al., 2009) (Veenstra, Dey, & Herrin, 2008) (Cheewaprakobkit, 2013) (Li, et al. 2013), and in other universities in Peru.

References

Baradwaj, B. K., & Pal, S. (2012). Mining educational data to analyze students' performance. *arXiv preprint arXiv:1201.3417*.

Bayer, J., Bydzovská, H., Géryk, J., Obsivac, T., & Popelinsky, L. (2012). Predicting Drop-Out from Social Behaviour of Students. *International Educational Data Mining Society*.

Byrne, M., & Flood, B. (2008). Examining the relationships among background variables and academic performance of first year accounting students at an Irish University. Journal of Accounting Education, 26(4), 202-212.

Calisir, F., Basak, E., & Comertoglu, S. (2016). Predicting academic performance of master's students in engineering management. *College Student Journal*, 50(4), 501-513.

Chalaris, M., Gritzalis, S., Maragoudakis, M., Sgouropoulou, C., & Tsolakidis, A. (2014). Improving quality of educational processes providing new knowledge using data mining techniques. *Procedia-Social and Behavioral Sciences*, *147*, 390-397.

Cheewaprakobkit, P. (2013). Study of Factors Analysis Affecting Academic Achievement of Undergraduate Students in International Program. *In Proceedings of the International MultiConference of Engineers and Computer Scientists* (Vol. 1, pp. 13-15).

Ecklund, A. P. (2013) Enhancing Incoming Male Student Retention: An Analysis of the Experiences of Persistence in Engineering.

Elakia, G., & Aarthi, N. J. (2014). Application of data mining in educational database for predicting behavioural patterns of the students. International Journal of Computer Science and Information Technologies (IJCSIT), 5(3), 4649-4652.

ElGamal, A.F. (2013). An Educational Data Mining Model for

Predicting Student Performance in Programming Course. International Journal of Computer Applications, 70(17).

Gray, G., McGuinness, C., & Owende, P. (2014). An application of classification models to predict learner progression in tertiary education. In *Advance Computing Conference (IACC), 2014 IEEE International* (pp. 549-554). IEEE.

Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques, Elsevier.

Honken, N. B., & Ralston, P. A. (2013). High-Achieving High School Students and Not So High-Achieving College Students A Look at Lack of Self-Control, Academic Ability, and Performance in College. Journal of Advanced Academics, 24(2), 108-124.

Lantz, B. (2013). Machine learning with R. Packt Publishing Ltd.

Li, Q., Swaminathan, H. and Tang, J. (2009), Development of a Classification System for Engineering Student Characteristics Affecting College Enrollment and Retention. Journal of Engineering Education, 98: 361–376. doi: 10.1002/j.2168-9830.2009.tb01033.x

Li, K. F., Rusk, D., & Song, F. (2013). Predicting student academic performance. In *Complex, Intelligent, and Software Intensive Systems (CISIS), 2013 Seventh International Conference on* (pp. 27-33). IEEE.

López Bonilla, J. M., López Bonilla, L. M., Serra, F., & Ribeiro, C. (2015). Relación entre actitudes hacia la actividad física y el deporte y rendimiento académico de los estudiantes universitarios españoles y portugueses. *Revista iberoamericana de psicología del ejercicio y el deporte*, *10*(2), 275-284.

Mishra T., Kumar D. & Gupta S. (2014) Mining Students' Data for Prediction Performance Fourth International Conference *on Advanced Computing & Communication Technologies*, Rohtak, pp. 255-262.

Ramesh, V., Parkavi, P., & Ramar, K. (2013). Predicting student performance: a statistical and data mining approach. *International journal of computer applications*, *63*(8).

Rodríguez, Á. P. A., & Arenas, D. A. M. (2016). Programas de intervención para Estudiantes Universitarios con bajo rendimiento académico. *Informes Psicológicos*, *16*(1), 13-34.

Romero, C., Ventura, S., Pechenizkiy, M., & Baker, R. S. (Eds.). (2011). *Handbook of educational data mining*. CRC Press.

Romero, C., & Ventura, S. (2013). Data mining in education. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3(1), 12-27.

Sembiring, S., Zarlis, M., Hartama, D., Ramliana, S., & Wani, E. (2011). Prediction of student academic performance by an application of data mining techniques. In International Conference on Management and Artificial Intelligence IPEDR (Vol. 6, pp. 110-114).

Sepehrian, F. (2012). Emotional Intelligence as a predictor of academic performance in university. *Journal of Educational Sciences and Psychology*, *2*(2).

Shahiri, A. M., & Husain, W. (2015). A review on predicting student's performance using data mining techniques. *Procedia Computer Science*, *72*, 414-422.

Strecht, P., Cruz, L., Soares, C., Merdes-Moreria, J. & Abren, R. (2015). A Comparative Study of Classification and Regression Algorithms for Modelling Students' Academic Performance. In 8th International Conference on Educational Data Mining, Madrid, Spain, 392-395.

Veenstra, C. P., Dey, E. L., & Herrin, G. D. (2008). Is Modeling of Freshman Engineering Success Different from Modeling of Non-Engineering Success?. Journal of Engineering Education, 97(4), 467-479.

York, T. T., Gibson, C., & Rankin, S. (2015). Defining and measuring academic success. *Practical Assessment, Research & Evaluation*, *20*(5), 2.