A system based on data science to predict student dropouts

March 31 2017

Credit: Universidad de Barcelona

There is currently a 30 percent student dropout rate in Europe, according to Education at a Glance (EAG). In Spain, these figures are between the 25 percent and 29 percent. In order to create a tool to assess and improve

academic performance and reduce these levels, a team of the Faculty of Mathematics and Computer Science has published an article in the journal *PLoS ONE* that presents a data analysis system to predict student dropout. This tool is based on machine learning techniques.

The study is signed by the researchers Laura Igual and Eloi Puertas, from the Department of Mathematics and Computer Science of the UB, together with Sergi Rovira, student of the bachelor degree in Computer Engineering of the UB. The aim is to develop a tool for the lecturers that gives recommendations for the students, and can assess the risk of student dropouts.

"Nowadays, the role of the tutor is more important than ever in order to prevent students from leaving the university and improve their academic performance. The research proposes a system based on objective data to take hidden information which is important for the students' academic data and therefore, to help teachers to offer their students a personal and proactive orientation," says Igual.

In this first stage, the objective of the research was to answer the question "is it possible to predict whether a student will continue the second year at university based on the results from the first academic year?" To conduct the analysis, the researchers used data from the first and second academic years in three bachelor degrees: mathematics, computer science and law. To do so, they applied five data science algorithms, the best of which has shown a precision of 82 percent. Both the algorithm and anonymous data are publicly published in *PLoS ONE*.

From statistics to data science

The previous studies on university dropouts in this field were focused on statistical models, based on a collection of data (usually through interviews) gathering information on the possible causes of study

dropout. Statistical models are based on hypotheses taken from the underlying problems. If students' performance factors change over time, the assumptions of a statistical model could be obsolete.

"However, machine learning techniques have a predictive use based on objective data, which makes them more adaptable to new data," says Igual. However, statistical systems are better at determining the reasons students leave their studies.

"But the predictive power of these tools is lower," says Laura Igual. Also, this new focus will allow the teaching staff to have "warnings" about students before registering, she notes.

This system also allows predicting the grades students will get in future courses, which would allow the teachers to give advice or orientation to students.

Within the teaching innovation project, "the following step is to analyze –from an educational perspective- how to use this tool, how to assess its impact and develop a computer application prototype," concludes the researcher.

Provided by University of Barcelona