

# A Case Study: Data Mining Applied to Student Enrollment

César Vialardi<sup>1</sup>, Jorge Chue<sup>1</sup>, Alfredo Barrientos<sup>1</sup>, Daniel Victoria<sup>1</sup>, Jhonny Estrella<sup>1</sup>, Juan Pablo Peche<sup>1</sup>  
and Álvaro Ortigosa

{cvialar, jchue, abarrien, dvictoria, jestrella, jpeche}@ulima.edu.pe, Alvaro.ortigosa@uam.es

<sup>1</sup>Facultad de Ingeniería de Sistemas, Universidad de Lima

<sup>2</sup>Escuela Politécnica Superior, Universidad Autónoma de Madrid

**Abstract.** One of the main problems faced by university students is deciding the right learning path based on available information such as courses, schedules and professors. In this context, this paper presents a recommender system based on data mining. This recommender system intends to create awareness of the difficulty and amount of workload entailed by a chosen set of courses. For the purpose of building the underlying model, this paper describes the generation of domain specific variables that are capable of representing students' past performance. The objective is to improve students' performance in general, by reducing the rate of misguided enrollment decisions.

## 1. Introduction

University Curricula allow students great flexibility and freedom of choice in terms of which and how many courses they can sign up for each term. The amount of workload and their ability to balance it successfully depends on these choices. Their results are also dependent on their own ability for a particular area of knowledge.

The main objective of this paper is to propose an enrollment recommender system to assist students in their decision making. The main contribution is the generation of two domain specific variables, namely the potential of student and the difficulty of courses.

A similar study was conducted by Al-Radaideh [1], in which he uses classification algorithms to evaluate the performance of students who studied the C++ course in Yarmouk University in 2005. To build a reliable classification model, it adopts the CRISP-DM methodology.

## 2. Domain Specific Variables

Domain specific metrics increase the representativeness of models. In this particular case we used two variables: the course difficulty and ability of a student towards a course; the latter is referred to as potential.

The course difficulty is represented by the average of the grades obtained by students. On the other hand, the potential is calculated per student for each course he may take. It is represented by the average of the grades a student has obtained in the prerequisites of a course and in previous attempts to pass the course; each grade is divided by the course difficulty.

## 3. Recommender System

The model that supports the recommender system was built using the Knowledge Discovery in Databases Methodology [2] using the C4.5 algorithm as classification engine [3]. It included the domain specific variables presented in Section 2 in order to improve its effectiveness. This system is integrated into the current Enrollment System.

During enrollment students choose a set of courses and obtains a forecast for each of them: PASS/FAIL. This enables them to make informed and conscious decisions hence indirectly improving their performance. (Figure 1) shows the sequence of this process in detail.

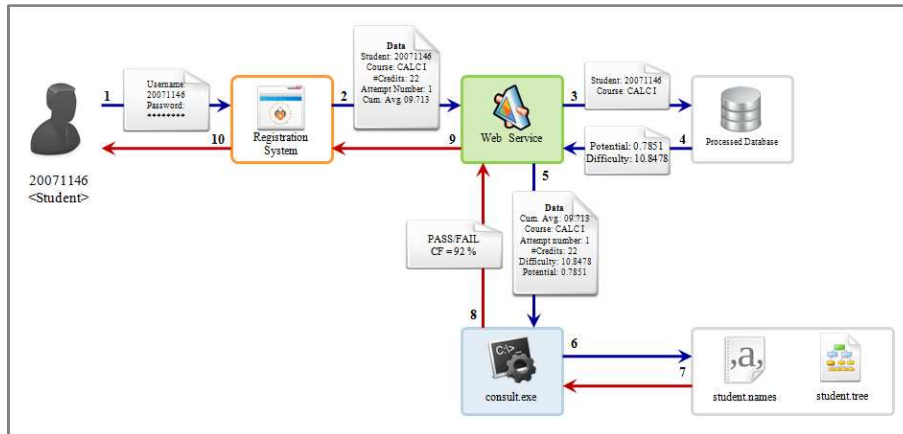


Figure 1. Recommendation Sequence Diagram

#### 4. Conclusions and Future Work

The main benefit of the proposed system is the awareness created in students about the enrollment process: they will be aware of the fact that they are not currently skilled enough to success in courses in a particular area; it will also help them appraise the difficulty of courses and realize the possibility of unbalanced workloads. In the long run, these positive effects will lower failure rate of students hence improving their learning process and their learning paths.

Future works should strive to improve data cleaning to remove noise from the model. The model itself could be enhanced by improving the C4.5 pruning method or by adding significant attributes.

#### Acknowledgement

This work has been funded by Spanish Ministry of Science and Education through the HADA projects (TIN2007-64718) and the Universidad de Lima through the IDIC (Research Institute).

#### References

- [1] Al-Radaideh, Qasem A., Al-Shawakfa, Emad M., and Al-Najjar, Mustafa. Mining Student Data using Decision Trees. *International Arab Conference on Information Technology*, 2006.
- [2] Fayyad, U., Piatesky-Shapiri, G., and Smyth, P. From Data mining to knowledge Discovery in Databases. *AAAI*, 1997, p. 37-54.
- [3] Quinlan, J.R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, California, USA, 1993.