USING DATA MINING TO PREDICT SECONDARY SCHOOL STUDENT PERFORMANCE

*Paulo Cortez and Alice Silva Dep. Information Systems/Algoritmi R&D Centre*

*University of Minho*

SUMMARY

The paper aims to investigate the predictability of study performance and which factors influence student achievements. In this case, Business Intelligence and Data Mining techniques are used to better understand the relevant features that affect student performance in secondary schools. Predicting student performance is an important tool as it improves the quality of education and simultaneously provides knowledge to better manage school resources.

The study uses real-world data retrieved from two Portuguese secondary schools as a reference for the model. The study focuses on two subjects, Mathematics and Portuguese, as they represent the core classes that provide the fundamental knowledge basis. Data is collected through school reports and questionnaires, where the two incomplete sources are combined to obtain the necessary information. The questionnaires allowed to identify multiple variables that could potentially affect educational success/failure. The variables used as inputs in the model include demographic, social, emotional, and school related attributes.

The DM algorithms used to construct the models are Decision Trees (DT), Random Forest (RF), Neural Networks (NN) and Support Vector Machines (SVM). These algorithms have different purposes and capabilities that are used to perform classification and regression goals. Moreover, the paper selected three classification approaches to model the grades: binary classification (pass/fail), classification through five levels (in a range from excellent to insufficient), and regression (based on a numeric output that ranges between zero and twenty). Also, the models are characterized by three different input setups that either include or exclude previous grades into the analysis. These inputs are based on evaluation that is divided over three periods, where the last result is equal to the final grade. As a matter of fact, the paper acknowledges that different studies confirm that past grades have a higher impact compared to demographic variables.

The model explicit different metrics that are useful for classification and regression, which both entail supervised learning. In the study, this process is assisted by R to conduct the experiments. For the classification, the evaluation is made through the Percentage of Correct Classifications (PCC), whereas the Root Mean Squared (RMSE) is mainly used for the regression. To identify an acceptable classifier, the PCC should be high and the regressor should be characterized by a low global error.

The results show that student evaluation have a high impact, ranging from 23% to 46%, as well as the number of past failures. In addition, the explanatory analysis reveals that there are several relevant factors that substantially affect the performance. These can range from demographic to social variables, including number of absences, parents’ job, or alcohol consumption. Therefore, we can affirm that the study on student’s performance can be considered as a guideline for both educators and students.