Educational Data Mining: An Application

of Regressors in Predicting School Dropout

Abstract. School dropout is one of the great challenges for the educational system. Educational data mining seeks to study and contribute with results that aim to hidden problems and find possible solutions. Considering its importance, this work aims to use two nonparametric techniques, Quantile Regression and Support Vector Regression, to predict the results of school dropout in the Brazilian scenario. The development of the work followed the phases of CRISPDM. The evaluation metric of the models is the mean of the absolute error. The results show more significant results for Support Vector Regression.

Keywords: Education Data Mining SVR NPQR Prediction School dropout

1 Introduction

School dropout represents a major problem that needs to be studied and wrestled. There is a governmental effort to reduce this rate, which interferes in the countrys educational development indexes, as well as directly impacts the personal and professional scope of Brazilian youth. The phenomenon of school dropout is seen as one of the greatest problems in any level of education [1], and knowing the reasons behind it, institutions can create mechanisms to reduce it. The National Institute for Educational Studies and Research Ansio Teixeira (INEP) is responsible for collecting and disseminating information about education in Brazil, at all stages of education, through assessments and indicators [2]. According to INEP [3], new researches show that 12.9% and 12.7% of the students enrolled in the first and second years of high school, respectively, evaded school as shown in the School Census between the years 2014 and 2015. The 9th year of elementary education has the third highest dropout rate 7.7%, followed by the 3rd year of high school with 6.8%. Considering all high school grades,

dropout sums to 11.2% of total students at this stage of education.

It is notable that combating evasion is still one of the great challenges for education, becoming a very relevant issue, which has been expanded and addressed in many studies. Some of the major research approaches to school evasion are prediction, clustering, relationship mining, discovery with models, and data processing for decision support.

One of the factors that helps the development of this work is the availability of the data related to education, increasing the applicability of Education Data Mining (EDM). By applying EDM, it is possible to effectively and accurately understand students, the role of the context in which learning takes place and other factors that influence the teachinglearning process [4]. This is of extreme interest to a wide variety of people, including educators, students, institutions, government, parents and the general public [5].

EDM focusing on finding answers to specific education questions related to learning processes, development of instructional materials, monitoring and predicting. From obtaining important information and behavior patterns it is possible to support certain pedagogical practices [4]. EDM can be interpreted as a process where the goal is not only to transform data into knowledge, but also to filter knowledge to help make decisions about how to modify the educational environment [6]. It is a field that exploits statistical, machinelearning, and data mining algorithms over the different types of educational data [7]. There are several lines of research in the area of education and many of them derived from the data mining area, such as predictive, grouping or association tasks [4].

Recently, EDM research papers have focused on predictive models to maximize student retention [8], enrollment prediction models based on admission data [9], student performance forecast [10] and school dropout [11]. An accurate predictive model can be used to gain insight into success and risk factors in relation to the educational environment.

In the field of prediction, techniques can be applied to discover structures or associations in data set and make predictions. Among them, regression models are emphasized, which are mathematical model and have as one of the objectives to predict the value of the dependent variable (Y) from the information coming from a set of independent variables (X) [12]. Therefore, this type of technique can estimate the educational benefits and problems.

In this way, this paper aims to predict school dropout by means of educational databases provided by INEP. The techniques used in the experiments were support vector regression (SVR) and nonparametric quantile regression (NPQR). The study followed the phases of the CrossIndustry Standard for Data Mining (CRISPDM) [13] model.

This paper is divided as it follows: Sect. 2 presents works related to the theme of this article; Sect. 3 the methodology, where the information about the databases and the techniques used in the development of this work are presented; Sect. 4 shows the results of the experiments performed; and, finally, Sect. 5 composes the conclusion of this work developed after the analysis of the results obtained at the end of the experiments.

2 Related Works

EDM has grown in recent years and the information provided by that can serve as a subsidy to improve education practices, as well as being an important tool to enable education qualification. More and more studies address the educational issue, especially school dropout.

In relation to the techniques of data mining applied to education, Machado et al. [14] made a bibliography review focused on to identify the papers which approach the problem of school dropout using data mining techniques. That study allowed to identify the main methods used in this subject, which are decision trees, neural networks, logistic regression and clustering algorithms.

In Martinho et al. [15], a Neural Network was used to present the prediction of the group of the students at risk of dropping out in higher education classroom courses. In Meedech et al. [11], it also addresses evasion and applies decision trees and rules induction models to discover student data knowledge. Such as in the research by Quadri and Kalyankar [16] that applies the decision tree technique to choose the best forecast and analysis.

Regarding distance education, Cambruzzi et al. [17] address the dropout rates observed in these courses which are very significant. The research presents a system of Learning Analysis developed to deal with the problem of abandonment in distance education courses in university education. Data visualization and text mining are used.

et al. [18] proposed the application of data mining techniques to predict school failure and school dropout. From actual data of Mexican students, they conducted experiments using induction rules and decision trees. Veitch [19] justifies the use of decision trees to predict school dropout because this method is designed to sift a set of predictor variables and successively divides a set of data into subgroups to improve the prediction (classification) of a target (dependent) variable. In addition to decision trees, the research by da Cunha et al. [20] implements grouping algorithms in the school dropout scenario.

In the work of Rodrigues et al. [21], the objective was to investigate the feasibility of using the linear regression model to obtain inferences in the initial stages of online courses, as a way of supporting decision making by teachers and managers. The results obtained demonstrated that it is possible to use the linear regression technique to obtain inferences with good accuracy rates.

However, few papers are still used in its technical experiments different techniques from those that were approached by Machado et al. [14] in his bibliography review. This is the case of quantile regression [22]. Quantile regression differs from the other regression types since it allows the use of several curves (or quantiles) to obtain a more complete view of the relationship between the studied variables. For nonparametric quantile regression model, the Gaussian kernel can be applied by adjusting the parameter bandwidth that controls the degree of smoothness of the estimated function.

Another technique for prediction is the Support Vector Machine (SVM) and it can also be used for regression, retaining all the key functions of the algorithm in which are, generating the maximum number of support vectors in order to

maximize the separation of data classes and to maximize the margins of these support vectors. The extension of the SVM technique for regression is the support vector regression (SVR). SVR algorithms perform well in nonlinear problems, such as time series [23]. The advantage of using SVR over other techniques because it focuses on finding the global optimal value. In addition, its model is easier to understand comparing to others well known techniques, such as artificial neural network.

In order to approach these presented works, the accomplishment of this research brings relevant contribution since the data mining for the educational environment is an extremely important subject and that needs to be studied deeply. The topic of evasion is much addressed in the works, so it is necessary to apply other techniques and verify the performance of the results. Thus, using the techniques NPQR and SVR the predictive analyzes of school dropout will be constructed.

3 Methodology

One of the most popular methodologies to increase the success of data mining processes is CRISPDM [13]. The methodology defines a nonrigid sequence of six phases which allows the construction and implementation of a mining model to be used in a real environment, helping business decisions [24]. Therefore, the development of this paper follows the phases of CRISPDM that is shown in Fig. 1 and described in the following subsections.

3.1 Business Understanding

The initial phase focuses on understanding project goals and requirements and converts this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.

A bibliographical research was carried out, checking all material already elaborated and related to educational scenarios, educational data mining and regression models. After that, the main variables involving the research subject were listed, which consists of predicting school dropout.