Educational Data Mining: An Application

of Regressors in Predicting School Dropout

**Abstract.** School dropout is one of the great challenges for the educa- tional 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 pre- dict the results of school dropout in the Brazilian scenario. The devel- opment of the work followed the phases of CRISP-DM. 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

*· · ·*

# Introduction

School dropout represents a major problem that needs to be studied and wres- tled. There is a governmental eﬀort to reduce this rate, which interferes in the country’s 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](#_bookmark10)], and knowing the reasons behind it, institutions can create mechanisms to reduce it. The National Institute for Educational Studies and Research An´ısio Teixeira (INEP) is responsible for collecting and disseminating information about educa- tion in Brazil, at all stages of education, through assessments and indicators [[2](#_bookmark11)]. According to INEP [[3](#_bookmark12)], new researches show that 12.9% and 12.7% of the stu- dents 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 edu- cation, becoming a very relevant issue, which has been expanded and addressed

Ⓧc Springer International Publishing AG, part of Springer Nature 2018 P. Perner (Ed.): MLDM 2018, LNAI 10935, pp. 246–257, 2018.

https://doi.org/10.1007/978-3-319-96133-0\_19

in many studies. Some of the major research approaches to school evasion are prediction, clustering, relationship mining, discovery with models, and data pro- cessing 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 teaching-learning process [[4](#_bookmark13)]. This is of extreme interest to a wide variety of people, including educators, students, institutions, government, parents and the general public [[5](#_bookmark14)].

EDM focusing on ﬁnding answers to speciﬁc education questions related to learning processes, development of instructional materials, monitoring and pre- dicting. From obtaining important information and behavior patterns it is pos- sible to support certain pedagogical practices [[4](#_bookmark13)]. 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](#_bookmark15)]. It is a ﬁeld that exploits statistical, machine-learning, and data mining algorithms over the diﬀerent types of educational data [[7](#_bookmark16)]. 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](#_bookmark13)].

Recently, EDM research papers have focused on predictive models to max- imize student retention [[8](#_bookmark17)], enrollment prediction models based on admission data [[9](#_bookmark18)], student performance forecast [[10](#_bookmark19)] and school dropout [[11](#_bookmark20)]. 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](#_bookmark21)]. 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 sup- port vector regression (SVR) and nonparametric quantile regression (NPQR). The study followed the phases of the Cross-Industry Standard for Data Mining (CRISP-DM) [[13](#_bookmark22)] model.

This paper is divided as it follows: Sect. [2](#_bookmark0) presents works related to the theme of this article; Sect. [3](#_bookmark1) the methodology, where the information about the databases and the techniques used in the development of this work are presented; Sect. [4](#_bookmark5) shows the results of the experiments performed; and, ﬁnally, Sect. [5](#_bookmark9) com- poses the conclusion of this work developed after the analysis of the results obtained at the end of the experiments.

# 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](#_bookmark23)] 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](#_bookmark24)], 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](#_bookmark20)], 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](#_bookmark25)] that applies the decision tree technique to choose the best forecast and analysis.

Regarding distance education, Cambruzzi et al. [[17](#_bookmark26)] 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.

M´arquez-Vera et al. [[18](#_bookmark27)] proposed the application of data mining techniques to predict school failure and school dropout. From actual data of Mexican stu- dents, they conducted experiments using induction rules and decision trees. Veitch [[19](#_bookmark28)] 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](#_bookmark29)] implements grouping algorithms in the school dropout scenario.

In the work of Rodrigues et al. [[21](#_bookmark30)], the objective was to investigate the feasibility of using the linear regression model to obtain inferences in the initial stages of on-line 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 tech- niques from those that were approached by Machado et al. [[14](#_bookmark23)] in his bibliog- raphy review. This is the case of quantile regression [[22](#_bookmark31)]. 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.

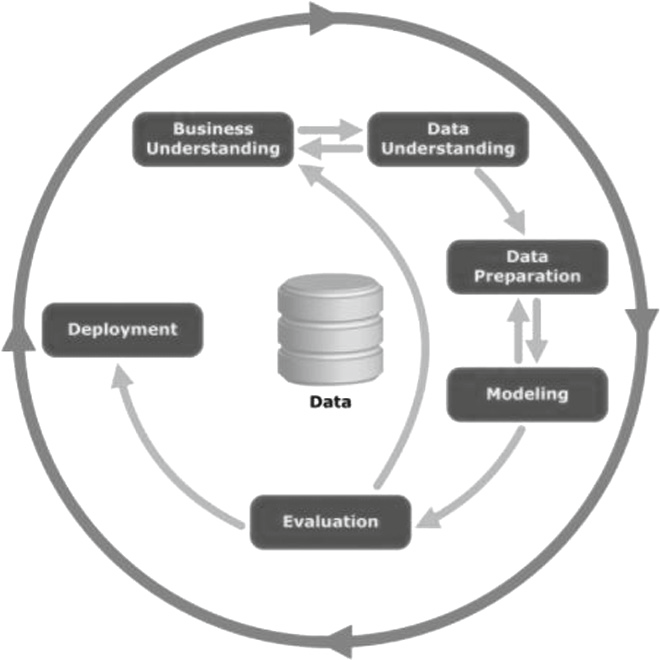
Another technique for prediction is the Support Vector Machine (SVM) and it can also be used for regression, retaining all the key functions of the algorithm in which are, generating the maximum number of support vectors in order to

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](#_bookmark32)]. 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.

# Methodology

One of the most popular methodologies to increase the success of data mining processes is CRISP-DM [[13](#_bookmark22)]. The methodology deﬁnes a non-rigid 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](#_bookmark33)]. Therefore, the development of this paper follows the phases of CRISP-DM that is shown in Fig. [1](#_bookmark2) and described in the following subsections.



**Fig. 1.** Steps of CRISP-DM [[13](#_bookmark22)].

## 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 elab- orated and related to educational scenarios, educational data mining and regres- sion models. After that, the main variables involving the research subject were listed, which consists of predicting school dropout.

## Data Understanding

The data understanding phase begins with initial data collection and proceeds with activities that allow to familiarize with the data, identify data quality prob- lems, ﬁrst discover insights about the data, and detect interesting subsets to form hypotheses about hidden information.

**Databases Description.** The databases used in this study were obtained from the open web portal of the INEP, referred to the microdata of the School Census and School Performance Rate, both from the year 2016.

The School Census is a survey of educational information of national scope carried out annually and made available by INEP. This research aims to carry out a survey on the basic education schools in the country. It covers beyond the stages of regular education, special education, youth and adult education and professional education. The information collected is classiﬁed into four major dimensions, schools, students, teachers and classes.

In this paper, it is used the schools dimension, where data are collected on infrastructure such as basic sanitation, electricity, science and computer labs, accessibility, as well as on stages and schooling modalities oﬀered, location and administrative dependency. The original basis of the 2016 School Census of the schools dimension has a total of 279,359 instances and 166 variables.

INEP also provides educational indicators which attribute statistical value to the quality of teaching, attending not only to students’ performance but also to the economic and social context in which schools are inserted. They con- sider information such as access, permanence and student learning. The educa- tional indicator used in this study refers to the database Performance Rate. The database has information on student approval, disapproval and dropout.

These data can be obtained in diﬀerent levels such as national, regional or school. Thus, the school level database is used since it is necessary to relate the values of this base to each school present in the base of the School Census. The original basis of this indicator has a total of 139,823 instances and 63 variables.

## Data Preparation

At this point, tasks include record and attribute selection as well as transforma- tion and data cleaning for modeling tools. In this phase, the following activities were carried out:

* Identiﬁcation of scenarios in the data.
* Selection of variables.
  + Checking for missing or blank values.
  + Data cleaning.
  + Data normalization.

Through each school’s identiﬁer, the school dropout rate present in the School Income database have been turned into a column of the School Census database. Then, the aspects of school infrastructure were correlated with the dropout rate. The data considered for study are of high school in the scope of the state of Pernambuco that contains 3,215 instances.

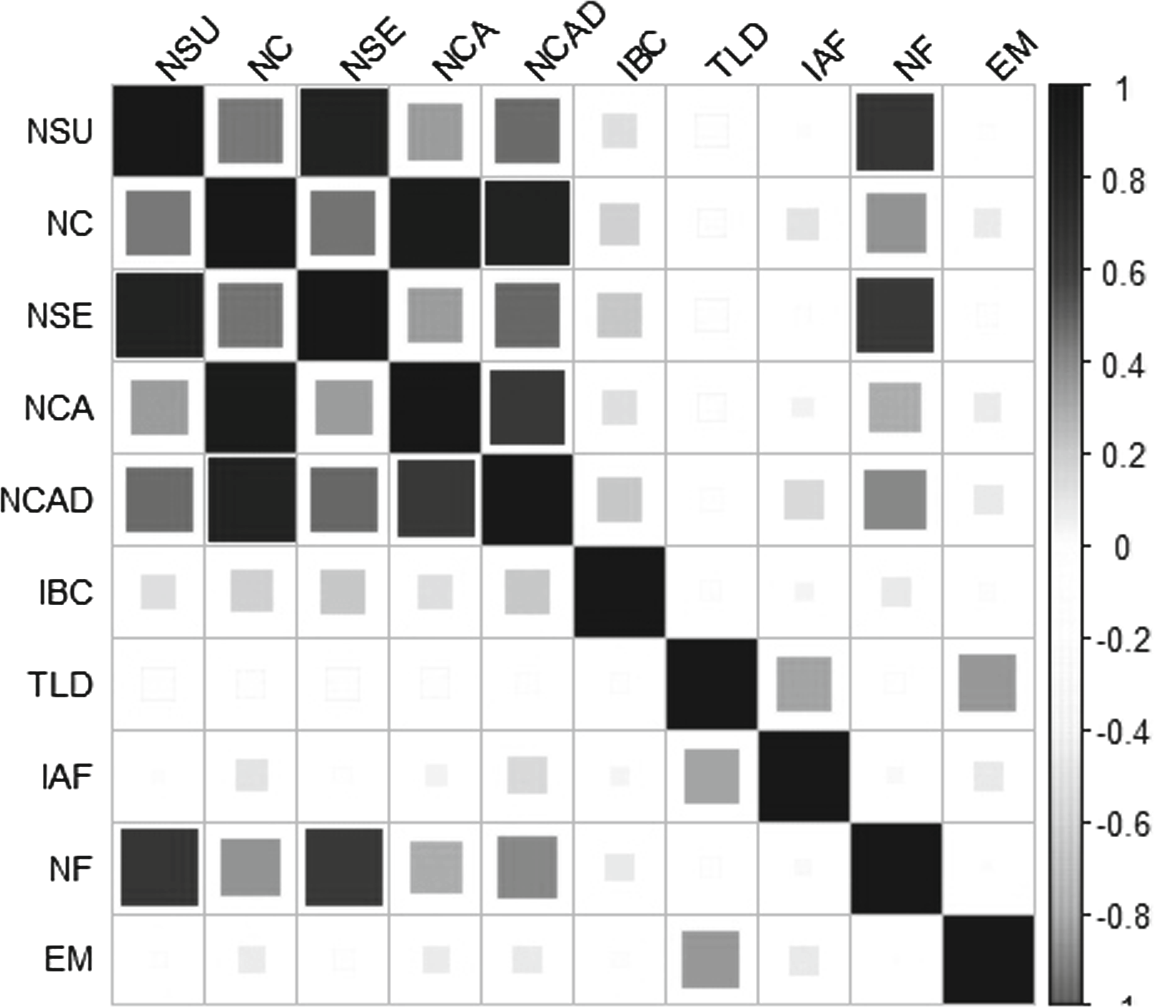
We used the Random Forest technique focusing on to obtain the degree of importance of each explanatory variable (X) in relation to the response variable (Y). Then, the 9 most important variables were selected for the model. At the end of this data preparation phase, the following data mapping was obtained, shown in the Table [1](#_bookmark3). After that, the correlation matrix of the selected variables was created, as shown in Fig. [2](#_bookmark4).

**Table 1.** Variables selected for study.

|  |  |  |
| --- | --- | --- |
| Ref | Variable | Description |
| NSU | NU SALAS UTILIZADAS | Total rooms used |
| NC | NU COMPUTADOR | Total computers |
| NSE | NU SALAS EXISTENTES | Total rooms that the school has |
| NCAD | NU COMP ADMINISTRATIVO | Total administrative computers |
| NCA | NU COMP ALUNO | Total computers for students |
| IBC | IN BANHEIRO CHUVEIRO | Bathroom presence indicator |
| TLD | TP LOCALIZACAO DIFERENCIADA | Type of school location |
| NF | NU FUNCIONARIOS | Total employees |
| IAF | IN AGUA FONTE RIO | Indicator of presence of water (river) |
| EM | EVASAO MEDIO | School dropout values |

By analyzing the Table [1](#_bookmark3), the variables related to the existence of computers in schools were identiﬁed and selected. The presence of computers in schools is a very important diﬀerential. This may indicate a better school structure as well as a concern of the school to develop technological teaching and improve digital access to information by students. Variables such as the existence of rooms, restrooms and the number of employees may also relate to the administrative capacity of the school, and consequently its structure. Such as the correlation matrix (Fig. [2](#_bookmark4)) was analyzed, there is a high degree of correlation between these variables (NSU, NC, NSE, NCA, NCAD and NF references, with coeﬃcients from 0.34 to 0.92), except for the presence of the variable of bathrooms in schools (IBC reference).

The location of the schools is also an important factor. Students who study far or in schools that are diﬃcult to access may be likely to be evaded. The last aspect to be analyzed is the issue of water supplied to school. The selected



**Fig. 2.** Matrix of correlation of the selected variables.

variable is related to the water supply through the river. This may be an indicator of schools that are not in areas near water stations, or further away from the urban environment. The correlation matrix shows a higher correlation between these two variables (TLD and IAF references, with coeﬃcient 0.32) compared to the others variables. The correlation matrix still informs that the variable with the highest correlation with school dropout (MS reference) is the location of the school (TLD), with coeﬃcient 0.35.

## Modeling

In this step, several modeling techniques were selected and applied, and their parameters were calibrated to optimal values. It represents the development of the models for the problem, based on the data that were already suitable to be used. The techniques used in this work consisted of NPQR and SVR.

**Nonparametric Quantile Regression:** To the use of the NPQR, the Gaus- sian Kernel was performed, and considered the quantile 0.5. This quantile was used because it presents better results in analyzes performed and it is considering the median of the data.

**Support Vector Regression:** The Grid Search approach was used to adjust SVR parameters. It was chosen: Radial Base Function Kernel; *ϵ* 1e-13, 1, 10; *γ* 1e0, 1, 1e-3, 1e-14; and *C* 1, 10, 100, 1000. These parameters are the most discussed in the literature.

## Evaluation

At this step of the project, the developed model is evaluated and the steps taken to create it are reviewed to be sure that it adequately achieves the deﬁned objectives.

One of the most used performance indexes in the prediction techniques is the calculation based on the prediction error. The performance index used in this work is the absolute mean error (MAE) that is a useful measure widely used in model evaluations [[25](#_bookmark34)].

## Deployment

All the knowledge obtained through the work of mining became subsidies for the development of strategies that solve the proposed problem. In this work, strategies will be listed for the scenario studied after all stages of CRISP-DM.

# Results of Experiments

In linear techniques, the predictors are combined in a linear fashion to model the eﬀect on the response. This linearity may be insuﬃcient to capture the data structure. However, there are techniques that allow a more ﬂexible regression modeling that combines the predictors in a nonparametric way. Thus, nonpara- metric regression techniques, NPQR and SVR were modeled seeking a response that best ﬁts the data.

The results obtained through the experiments were organized in a table and graph box-plot, in relation to the Median Absolute Error (MAE). The conﬁgura- tions of the techniques were described in the previous sections and performed 30 independent simulations for each experiment to form a signiﬁcant sample. There- fore, the mean of the executions of the techniques was extracted as it shows on Table [2](#_bookmark6).

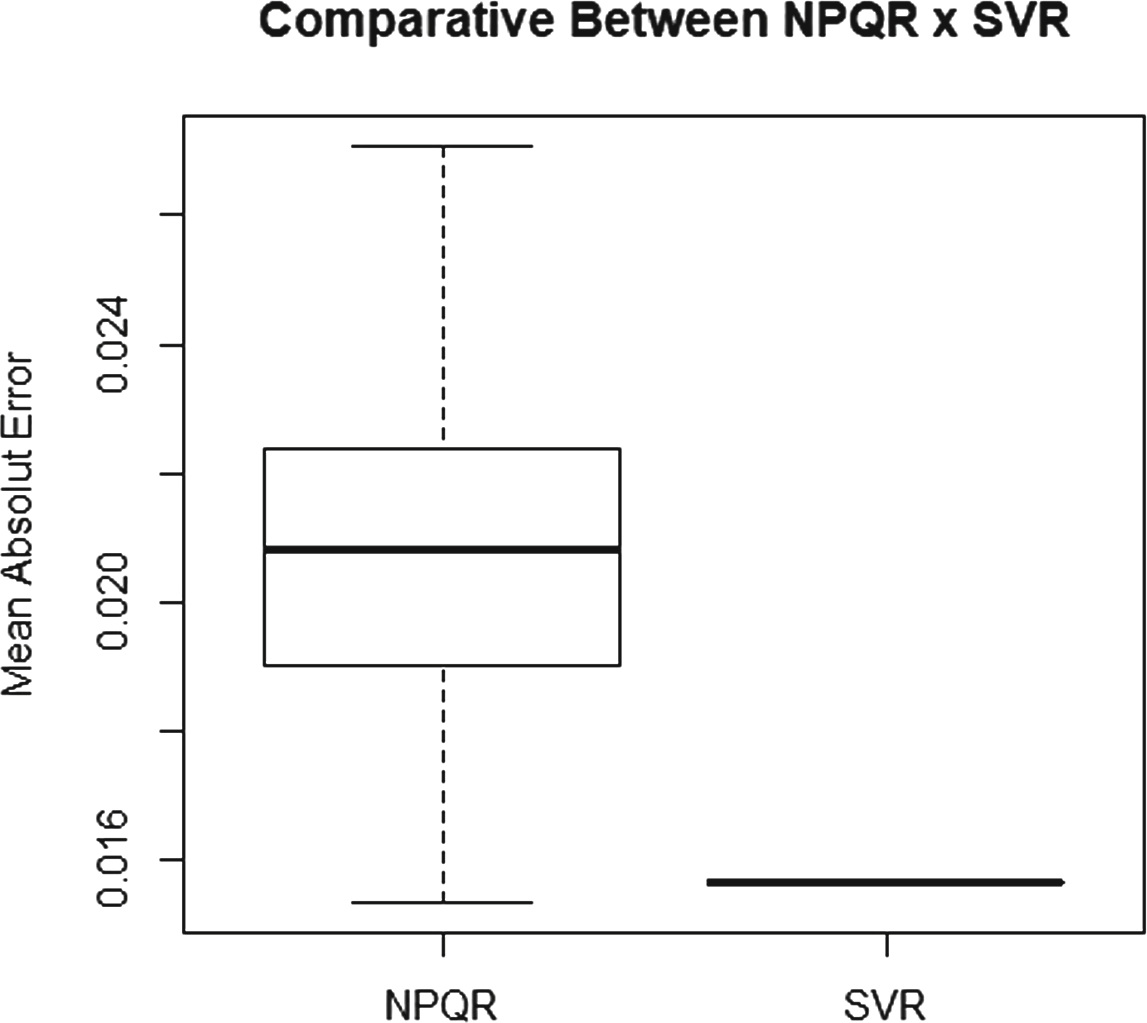
It can be veriﬁed that the SVR obtained smaller error comparing to NPQR. In analysis of the box-plot graph shown in the Fig. [3](#_bookmark7), the results of the techniques were compared side by side. One can note that SVR obtained smaller median in comparison to NPQR and the level of variability of the data contained in the NPQR sample is greater. This result shows that, indeed, SVR was able to ﬁnd the best combination of its parameters in all executions for predicting the school output.

Both techniques used in this work use a Kernel function to perform the esti- mation of the model. However, the results achieved by each technique are signiﬁ- cantly diﬀerent. The SVR is able to achieve better results because it searches for the optimal value. Another analysis is that the result of NPQR depends directly on how well estimated the bandwidth value is. However, estimating this value is not an easy task. This may cause lower NPQR performance.

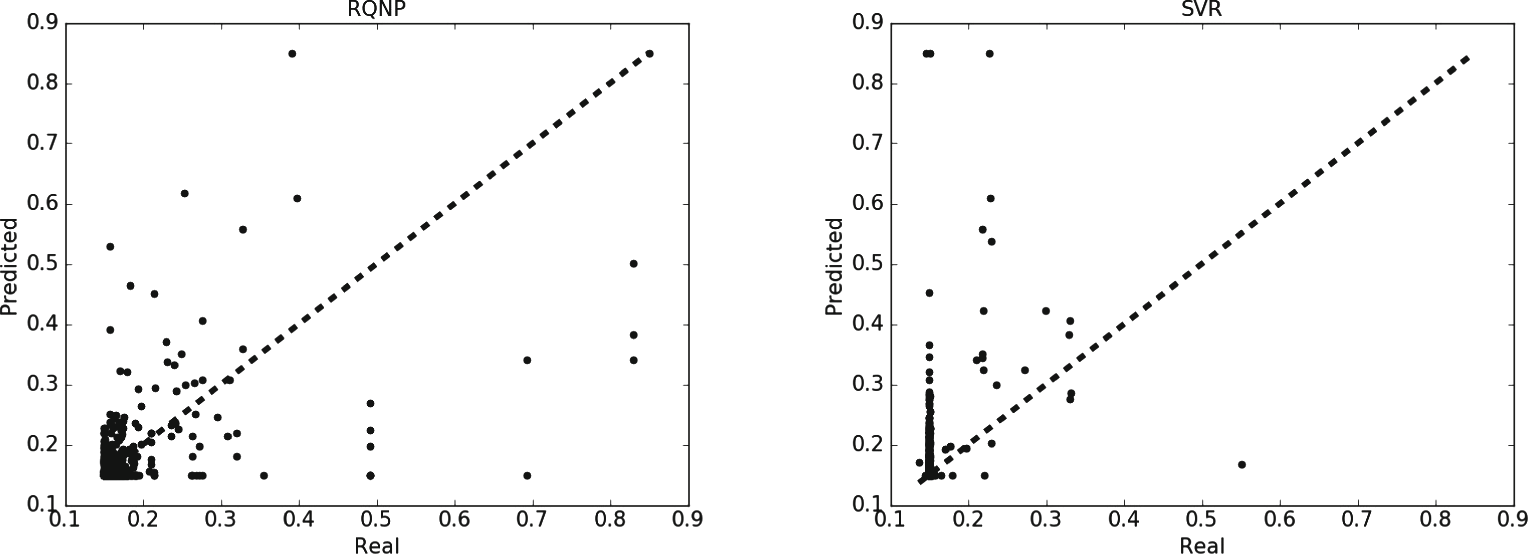
Through these analyzes, the SVR is doing a better work than the NPQR model at predicting school dropout. The Fig. [4](#_bookmark8) shows the visual plot for the

**Table 2.** Results NPQR X SVR

|  |  |  |
| --- | --- | --- |
| Technique | Mean error | SD error |
| SVR | 0.015665 | 0.0 |
| NPQR | 0.0223757 | 0.00324569 |



**Fig. 3.** Boxplot NPQR x SVR.



(a) NPQR (b) SVR

**Fig. 4.** Plot Predicted x Real of techniques.

performance of the techniques on the test set with predicted and real values. By visually inspecting the plot we can see that the predictions made by the SVR model are, in general, more concentrated around the line with less outliers than

those made by the NPQR model. In this way, we can conclude that SVR made a more accurate prediction.

With this result, it can be observed that nonparametric regression techniques can be applied in education. The power of adjustments and ﬂexibility to the data justiﬁes the applicability of this type of modeling when the parametric techniques are insuﬃcient. The educational area has several challenges. Using techniques associated with EDM to better estimate the variables related to these scenarios brings a great gain to the literature and to those interested in the area, such as students, educators and government, for example. Despite its advantages, in the literary this type of modeling is still little approached in educational problems.

# Conclusion

EDM enables the prior identiﬁcation of aspects that may need improvements and better investments, thus improving teaching, learning and mitigating problems such as school dropout. In this scenario, this study seeks to contribute to the area of Education in terms of understanding and explaining the characteristics of schools that may be related to school dropout. Public databases provided by INEP were used in this paper.

Through the Random Forest technique the variables that are most related to the problem of school dropout in high school were identiﬁed, according to the database provided. The existence of computers in schools, the number of employ- ees, rooms and the location of schools were highlighted as the most important for the model of prediction of evasion and were selected for the study.

In addition, it consisted of investigating new diﬀerent prediction techniques from those most used in the literature for EDM. Thus, experiments with the SVR and NPQR techniques were used. The NPQR was used to allow the use of several curves (quantiles) to obtain a more complete view of the relationship between the studied variables. The advantage of using SVR is that it searches for the optimal value and the model is easier to understand. The development of this work followed the phases of CRISP-DM.

The results of the experiments indicated that SVR obtained better results in comparison to NPQR. This means that SVR was able to minimize the prediction error for the school dropout problem.

The nonparametric techniques used bring a more ﬂexible regression model- ing, looking for a response that best matches the data. These techniques bring a gain to the educational area towards a better response when parametric tech- niques are insuﬃcient. With that being said, predicting variables related to teaching-learning can make research more accurate. These results made it pos- sible to investigate aspects in which the public policies for the scenario studied can be better applied.

In this way, for future work, it is sought to reﬁne forecasting techniques, as well as explore other educational aspects, expand the scope of study to other levels of education and to other scenarios such as regional or national.