Clustering of secondary school students in Portugal

D A Petrusevich

Russian Technological University (MIREA), Prospekt Vernadskogo, 78, Moscow, 119454, Russia

E-mail: petrdenis@mail.ru

Abstract. The dataset about the secondary schools in Portugal has been handled in the paper. Nowadays data analysis and mathematical statistics methods allow researchers and staff of universities to understand hidden dependencies in the data about students. In the original data competition for which the handled dataset was presented the main goal was to explain the final exams grades by means of social and behavioral parameters of a person. In the paper this question is researched in a new way. The clustering technique allows dividing students into a few groups. Mathematical models of the final grade are special for each cluster. Thus, models achieve some kind of individuality saving generality. Comparison of results of models constructed for the whole dataset and for each cluster has been prepared. Such data analysis technique can be implemented to handle another datasets with different set of features. Obtaining results of data analysis the staff is able to make conclusions on individual way of dealing with every cluster or students and some clusters can be analyzed in individual manner.

1. Introduction

In this paper the Students’ Performance dataset [1] has been explored by means of the machine learning algorithms [2]. Nowadays this approach is implemented in various domains of knowledge and sometimes it allows to find and analyze hidden dependencies in data. The handled dataset has been designed inside of the Students’ Performance data competition. The main goal of the contest was to understand whether there are some social or behavioral features that are connected with students’ grades. The best solution should be to find some features that can be influenced in order to increase students’ grades.

In the original research [1] prepared by the authors of the dataset the task of the students’ performance prediction has been solved. In this research the clustering technique is implemented to divide students into different groups that can be handled individually. Regression models are constructed for the whole dataset and for each cluster separately.

2. The dataset structure

The Portuguese language final grade is the main parameter that shows the quality of educational process in this dataset. There are three types of features. The first group consists of binary features: activities (extra-curricular activities), address (urban or rural), famsup (is there some educational support from family?), higher (does this student want to take higher education?), internet (has this student got access to internet at home?), nursery (attended nursery school) paid (has this student got additional paid classes?), Pstatus (do parents live together or apart?), romantic (has this student got a relationship?), school (Gabriel Pereira or Mousinho da Silveira school), schoolsup (is there extra educational support?).
sex. The second type of features consists of factor values with integer parameters in special diapason: age is between 15 and 22 years; famrel (quality of relationships in the family), Medu and Fedu (level of parent’s education) Mjob and Fjob (type of parent’s job), freetime and goout (how much free time has he or she got? How often does this student go out?), Dalc and Walc (daily and weekend level of alcohol consumption), health status have got 5 possible levels. Variables reason (reason to choose certain school), traveltime and studytime (how much time does student spend on way to school, doing homework?), failures (number of past class failures) have got 4 levels. Student’s guardian (mother, father, other) is explained with the guardian parameter. Features absences and G1, G2, G3 grades can also be treated as factor variables but there are much more levels. Grades belong to diapason between 0 and 20.

One can notice that the majority of the features are factor variables that have got limited number of possible values. In order to construct linear regression models all such parameters should be transformed into lists of dummy variables (using one-hot encoding) [3, 4]. At the same time some features of this nature can be considered as usual numbers: grade values. There are three marks. Two of them (G1, G2) are intermediate ones and the last one (G3) is the final grade. It depends on intermediate values. That’s why in the most of computational experiments only the G3 mark is used.

In the dataset performance of students of two schools is presented: Gabriel Pereira and Mousinho da Silveira. The age of the students is between 15 and 22 years.

Number of absences is transformed according to expression (1):
\[ x' = \frac{x - \mu}{\sigma}. \]

Here \( x \) is a handled value, \( x' \) is a transformed one, \( \mu \) is its mean value of \( x \) and \( \sigma \) is its standard deviation [3, 4].

3. Experiments
The investigated dataset [1] has been clustered into five different groups. Number of clusters has been determined by means of agglomerative clustering techniques [5] with use of the Ward’s method and the Euclidean distance (figure 1). Number of clusters is obtained with horizontal line setting at some height. Number of intersections with such line is equal to number of clusters. Here five clusters are separated very well from each other but one of them is quite small by size. Other clustering approaches [6] could lead to different structure of clusters. But usually there’s a lot of common features.

---

**Figure 1.** Agglomerative clustering of the Students’ Performance dataset (five clusters are used).
3.1. Clustering of the dataset
At the figure 1 it’s clearly seen that the data can be divided into two large groups. To understand this division mean values of all parameters have been counted for the both clusters. The first one contains the records of students with lower grades. The features which differ by more than 10 percent are shown below and this difference is analyzed. Smaller differences can be considered as random spikes. Number of clusters in any kind of cluster analysis is determined heuristically [6]. Here four large clusters and one small cluster (by number of objects) are situated far from each other (the Euclidean metrics is used). More thorough division can impede generalization of the constructed models.

In the table 1 one can see simple structure of the clusters. To explain it mean value of grades G3 are used. The process of division is hierarchical. Thus, "parent clusters" are divided into parts that can be considered as “child clusters”. This action can be repeated recurrently.

| Name of the cluster | Quantity of objects | Mean value of the G3 grade | Parent cluster |
|---------------------|---------------------|----------------------------|----------------|
| C1                  | 368                 | 9.30                       | dataset        |
| C2                  | 281                 | 13.89                      | dataset        |
| C11                 | 16                  | 0.06                       | C1             |
| C12                 | 265                 | 9.86                       | C1             |
| C121                | 48                  | 7.71                       | C11            |
| C122                | 217                 | 10.33                      | C11            |
| C21                 | 190                 | 12.38                      | C2             |
| C22                 | 178                 | 15.51                      | C2             |

The first cluster C1 contains students with lower G3 grade. There are 336 students of overall 649 ones. The majority of students in this cluster belong to the Mousinho da Silveira school. Students of this group have got more absences. One can conclude that they are worse students by academic activities. Among them there are more males. More often they live in rural areas than in urban ones. Level of their parents’ education is lower. There are more students in the second cluster whose parents have got higher education. Mothers of the students in the first cluster work at home more often.

The distance from school to home is higher in the 1st cluster. But difference isn’t large. In the first cluster there are more students that have to pay 15 – 30 minutes. In the second one there are students that pay 15 minutes or less to get to school.

Low weekly study time is met more often in the first cluster. The majority of the students in the 2nd cluster C2 have got study time between 5 and 10 hours.

Inside of another data science competition [1] the alcohol consumption by students has been analyzed. The dataset has been the same. The cases of high alcohol consumption are more often in the 1st cluster. This is true for the both kind of days: weekends and usual work week days. The majority of cases of very good health are presented in the 2nd cluster.

The same algorithm of agglomerative clustering has been implemented to the both clusters. If one considers the 1st cluster C1 with lower grades it’s divided into three parts: C121, C122, C11. There’s a small cluster C11 that contains 16 records with extremely low grades. There are no general conclusions on this cluster so the other part C12 of the 1st cluster allows to make two more clusters: C121, C122.

The C122 cluster with better academic results is going to be discussed. The number of absences is less, the grades are higher than in the C121 cluster. In this cluster the majority of students passes their exams and in the other cluster usually there are 1 or 2 past class failures. The family educational support is higher and students tend to get higher education though in common their parents’ level of education is lower. More students answer that their health level is very good.

The cluster C2 with higher grades can also be divided into two parts. There are fewer absences in the part with higher grades C22. So, this part can be explained as the one having better academic results. At
the same time their parents have got higher education more often. So, these students tend to get higher education in social way. The other part gets extra educational support more often.

3.2. Linear regressions in the constructed clusters

In the clusters that are shown above there have been constructed linear regression models [3, 4] in order to explain behavior of the G3 grade parameter by means of all other values in the dataset. First of all there’s a full model in which G3 variable is linear combination of all other parameters. The next step is construction of a reduced model. All insignificant terms are removed from this model. If some parameters become insignificant after such reduction they are also removed.

In the cluster C121 the determination coefficient $R^2$ value [3, 4] of the full model is 89.7% and the adjusted coefficient $R^2_{adj} = 22.6\%$. It means that there’s a lot of other factors that aren’t included into this dataset which could explain behaviour of students. All terms in the regression are insignificant though the determination coefficient value is high and one can conclude that parameters explain dynamics of the G3 value well.

In the cluster C122 the determination coefficient $R^2$ value of the full model is 41.5% and the adjusted coefficient $R^2_{adj} = 17.5\%$. It means that there is a lot of other factors that aren’t included into this dataset which could explain behaviour of students. But still there are significant terms that need to be considered thoroughly.

The insignificant terms have been removed from the model. The expression (2) shows the result of this process:

$$G_3 = 10.20 + 0.72 \text{paid} + 0.65 \text{failures0} - 0.47 \text{Dalc3}.$$  

(2)

Here $\text{failures0}$ is equal to 1 for students that have got zero failures and $\text{Dalc3} = 1$ for students with medium daily alcohol consumption. So, alcohol consumption decreases the grade and students without failed exams usually have got higher grades.

$R^2$ value of the reduced model is about 11.5%. Thus, there are factors which aren’t reflected in this model.

In the cluster C21 the determination coefficient $R^2$ value of the full model is 51.6% and the adjusted coefficient $R^2_{adj} = 22.6\%$. These values can be considered as quite small but the reduced model can be constructed. It’s shown in the expression (3):

$$G_3 = 13.31 - 0.42 \text{school sup} + 0.39 \text{age3} + 0.53 \text{freetime2} + 0.36 \text{freetime4} + 0.72 \text{freetime5} - 0.35 \text{health3}.$$  

(3)

Here $\text{school sup} = 1$ for students with extra educational support, $\text{age3}$ is 1 for students of the 3rd course, $\text{freetime2} = 1$ for students with less quantity of free time than average values, $\text{freetime4} = 1$ or $\text{freetime5} = 1$ means that there’s a lot of free time after lessons. $\text{health3} = 1$ if current health status is normal. In this model $R^2 = 15.5\%$.

In the cluster C22 the determination coefficient $R^2$ value of the full model is 68.3% and the adjusted coefficient $R^2_{adj} = 18.7\%$ and also there are a few significant terms that need to be considered thoroughly.

The reduced model is shown in the expression (4):

$$G_3 = 16.36 \text{failures0} + 0.48 \text{ Fedu2} - 0.44 \text{romantic}.$$  

(4)

Here $\text{failures0} = 1$ for students that with zero past class failures, $\text{ Fedu2} = 1$ for students whose fathers have got 5th – 9th grade education, $\text{romantic} = 1$ for students with relationships. Thus, relationships decrease final grades but in common this effect “costs” about 1 point. Also, if there are no past class failures the grade should be higher. In this model $R^2 = 7.9\%$.

Full model constructed for the whole dataset has got $R^2 = 43.3\%$, $R^2_{adj} = 35.9\%$. The reduced one has got values of determination coefficients: $R^2 = 32.0\%$, $R^2_{adj} = 31.2\%$. Thus, full models in clusters are
better by value of the coefficient of determination. But the reduced ones look worse: values of the R² coefficient is better in the model constructed for the whole dataset.

Overall results of the linear regression model for the whole dataset and models for each cluster are presented in the table 2. The models constructed for special clusters have got better R² values everywhere except cluster C122. At the same time the best reduced model has been constructed for the whole dataset. In this task there are too few records to construct good models. Low values of determination coefficients let us conclude that there are unobservable parameters or that students’ grades are explained with these parameters in a bad way.

### Table 2. Comparison of linear regressions constructed for each cluster and for the whole dataset.

| Name of the cluster | R², R²adj of the full model, % | R² of the reduced model, % |
|---------------------|--------------------------------|--------------------------|
| The whole dataset   | 43.3, 35.9                     | 32.0                     |
| C121                | 89.7, 22.6                     | –                        |
| C122                | 41.5, 17.5                     | 11.5                     |
| C21                 | 51.6, 22.6                     | 15.5                     |
| C22                 | 68.4, 18.7                     | 7.9                      |

#### 3.3. Decision tree regressions in the constructed clusters

Another technique that unites clustering or classification and linear regressions construction in each cluster or class is the decision tree regressor algorithm [7]. It uses the same structure and logic as implemented in the decision tree algorithms (CART) but in each node there’s some cluster of handled data and for this cluster there’s some regression model. The models have been trained at the set of 70% of records of a cluster or of the whole dataset and they’ve been tested at special set containing 30% of records. Size of training set is given at the 2nd column. Square root of the mean squared error (RMSE) [3, 4] is the main parameter used to measure quality of these models.

### Table 3. Comparison of root mean square error of the decision tree regressor models constructed for each cluster and for the whole dataset.

| Name of the cluster | Size of the test set | RMSE   |
|---------------------|----------------------|--------|
| The whole dataset   | 195                  | 3.15   |
| C121                | 14                   | 1.39   |
| C122                | 65                   | 1.24   |
| C21                 | 57                   | 1.03   |
| C22                 | 53                   | 1.63   |

Looking at the table 3 one can see that error values of the models constructed in clusters are less than error of the model trained with use of the whole dataset. Though relative values are better for the clusters C21, C122 containing more records than C121 and C22.

It’s difficult to make some conclusions using tree structures but the feature importance values can be useful. These characteristics show how important certain feature is while constructing decision tree regression. The most important features are shown in the table 4. It should be mentioned that there’s a lot of dummy variables. Here only the source features that have been splitted into a set of variables are shown.

### Table 4. The main features used in order to construct decision tree regressors in various clusters.

| Name of the cluster | Main features to construct decision tree regressors |
|---------------------|---------------------------------------------------|
| C121                | mother’s level of education (Medu), studytime, nursery |
| C122                | school, mother’s job (Mjobhealth), level of health, number of failures at exams |
| C21                 | sex, father’s job (Fjobservices), freetime, daily alcohol consumption (Dalc) |
Thus, in the clusters C121 and C22 variables on mother’s level of education (or parents’ level of education in common) and studytime are important. Parents’ type of job and student’s level of health or alcohol consumption are important features in the other clusters.

4. Conclusion
The secondary schools in Portugal dataset [1] has been investigated in this paper. The main goal of the original data contest for which this dataset has been designed was to predict the final grade of a student. The next question that needs an answer is to understand dependencies of the grades on the social and behavioral characteristics of students. Here the clustering of the dataset has been performed. Thus, data of students in each cluster can be analyzed separately. It’s comfortable to distinguish clusters by mean value of final grades. There are four main clusters and a small one.

There have been constructed data analysis models for grades prediction for each cluster. Thus, important features have been obtained. The linear regression models have got better \( R^2 \) values in clusters than the model trained with the whole dataset. Also decision tree regressor models have been constructed.

Determination coefficient values of linear regressions constructed for clusters are better than the coefficient value for the whole dataset for the majority of cases. Though the best reduced model (insignificant terms have been removed) is made for the whole dataset. It’s explained with low number of records in the dataset.

Decision tree regressors have also been constructed for each cluster. In the table 4 the most important features for automatic decision tree construction have been presented. It’s possible to conclude that there are social features logically connected with quality of study process. At the same time these features are quite individual. Models designed for clusters behave better than models made for the whole dataset.

The clustering itself can be an interesting result for staff. Students of each cluster can be handled by staff in its own way. It’s also possible to make groups of students balanced between clusters or to include into some groups only students from certain cluster.

Classification into clusters divided with limit values of grades doesn’t work well in this task. Here and in the original research [1] the \( F_1 \) value is about 60 – 80 % for various classification algorithms. It’s possible to interpret this result in the same way like it’s done for the linear regression models. Parameters of the dataset [1] are important but very often they’re insignificant in the models. Thus, students’ performance modeling and help from staff should be based on an individual approach.

Though in the paper [8] this dataset has been handled with special data analysis methods and the 80 – 90 % level of accuracy has been achieved it should be mentioned that this accuracy is obtained in classification into two groups task. The groups consist of students with marks that are higher or lower than mean grade. But the classifier is based on the features that can be insignificant. Strict results in this dataset handling may cause overfitting of classifiers. At the same time there are special wrapper methods to construct new datasets which are comfortable to build classification algorithms. In the present paper there’s a lot of attention paid to the clustering of students that seems more useful for university staff result than construction of classifiers with high-level accuracy.

One cannot say that there’s a lot of datasets on students’ performance in secondary or higher schools. Dependencies in parameters and features of students and universities are very useful and interesting values to analyze. Thus, new approaches in education can be developed and implemented. One of such examples is a blended learning system implementation [9]. These systems are developing rapidly. The Covid-19 pandemic makes university staff pay attention to them and to analysis of students’ performance at various stages of their implementation.

References
[1] Cortez P and Silva P 2008 Using Data Mining to Predict Secondary School Student Performance
Proc of the 5th Future Business Technology Conf (FUBUTEC 2008) EUROSI pp 5-12

[2] Sigov A S and Andrianova E G and Zhukov D O and Zykov S V and Tarasov I E 2019 Quantum informatics: Overview of the main achievements Rossiyskiy tekhnologicheskiy zhurnal (Russian Technological Journal) 7(1) pp 5-37 doi: 10.32362/2500-316X-2019-7-1-5-37

[3] James G and Witten D and Hastie T and Tibshirani R 2015 An introduction to statistical learning with applications in R (Springer-Verlag New York) 426 p. doi: 10.1007/978-1-4614-7138-7

[4] Hastie T and Tibshirani R and Friedman J 2009 The elements of statistical learning (Springer-Verlag New York) 533 p

[5] Reddy M and Makara V and Satish R U V N 2017 Divisive Hierarchical Clustering with K-means and Agglomerative Hierarchical Clustering Int J of Comp Science Trends and Tech (IJCST) 5(5) p 5-11

[6] Anfyorov M A 2019 Genetic clustering algorithm Rossiyskiy tekhnologicheskiy zhurnal (Russian Technological Journal) 7(6) pp 134-50 (In Russ.) doi: 10.32362/2500-316X-2019-7-6-134-150

[7] Breiman L and Freidman J H and Olshen R A and Stone C J 1984 Classification And Regression Trees (Monterey CA Wadsworth & Brooks/Cole Advanced Books & Software) 358 p. ISBN: 0534980548

[8] Unal F 2020 Data Mining for Students Performance Prediction in Education IntechOpen doi: 10.5772/intechopen.91449

[9] Lomonosova N V and Zolkina A V 2018 Digital learning resources: Enhancing efficiency within blended higher education Novosibirsk State Pedagogical University Bulletin 8(6) 121-37 doi: 10.15293/2226-3365.1806.08