Machine Learning applied to student drop-out problem

Ricardo Manhães Savii†, Zuleika Stefânia Sabino Roque, Juliana Garcia Cespedes
Instituto de Ciência e Tecnologia, Universidade Federal de São Paulo
S. J. dos Campos, SP, Brazil
Email: †ricardo.manhaes@unifesp.br

***Abstract*—This study is a product of a proposition of an interdisciplinary work dealing with debates among Human Rights, Multiculturalism in relationship with the social build of knowledge, science and technology. Its contribution and relevance is for Computer Science applied to Education by building a framework of data analysis aligned with the actual scenario of the Brazilian university education. On the last decade the number of places in universities doubled, and now the focus is to support these new students to maintain their positions as students and successfully complete a college degree. As education is a social right public politics are focusing that minorities and historically exploited parts of the population are given support. This demands a constant sight and efficient planning to ensure that this public enters, and are given the opportunity to complete their degrees. In public universities entire departments are created to support students in need, but students in difficult situation might have difficulties to contact this teams. And due to limited resources and personnel some students might not have the chance to get the help, promised by laws. This work presents some techniques of Machine Learning, like visualization that can help to identify, in matter of seconds, students with risk of drop-out, among thousands of other students, and statistics that can provide information for student support politicies and guidelines. As a final product of this work the intention is to create an open software repository containing the entirety of code, enabling institutions and support teams to use it on their own realities. For this study it was utilized a dataset from the Universidade Federal de São Paulo (UNIFESP) detailed later in this work.**

***Index Terms*—Student dropouts, Affirmative Actions, Machine Learning.**

I. Introduction

The 2011 education report of the Organization for Economic Co-operation and Development (OECD) shows that only 11% of Brazilian population between 25 and 64 years old hold a college degree. Between 1990 and 2010 there was a great raise on numbers of entry places for Universities in Brazil. So that, according to Ministério da Educação (MEC), there were 1,500,000 enrollments in 1991, in 2007 this number reached 5 millions and 6,739,689 enrollments were registered in 2011, source Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep).

The main problem to study this field is the lack of data [1]. And the evasion rate is considered an essential quality indicator of an University [2]. It cant signal institution dificulties or even social and regional problems, but to reach a diagnosis more in-depth studies are needed [3]–[5] and a proactive stance to obtain this data an analysis. A study realized with Inep[1], data from 2000 to 2005 [4] have shown that 2,165 higher education institutions, being 89.3% private, received 4,453,156 students for 20,407 undergraduate courses. And that during that period there was a 22% annual evasion rate[2]. It also showed a noticeable difference in the annual average rate of evasion between private and public institutions, where, respectively, they were 26% and 12% of dropouts. Some other points that also drew attention in the study of *Silva Filho et al.* [4] were: 1) the Southeast Region has about half of Brazil's higher education students. 2) Courses in Science, Mathematics and Computing have the highest dropout rates, 28%, considerably above the national average. 3) the Mathematics Course leads the list of highest rates with 44% dropout in 2005.

Machine Learning can be a useful tool for extracting important information when defining **student permanency strategies**. Such techniques that can detect patterns and react accordingly. For Stanford University [3], Machine Learning is a field that focuses on Induction and other types of algorithms that can be used to "learn". It is already used in several fields, such as: fraud detection, web search results, real-time marketing applications and websites, analysis of feelings in texts, credit and offers level, equipment failure prediction, Pricing models, intranet network intrusion detection, pattern recognition in images, spam filters and more.

Therefore, the main objective of this work is to identify techniques capable of obtaining good results in the analysis and identification of students at risk of evasion. More specifically, well-known computational techniques such as Gaussian Naive Bayes (GNB), Decision Trees (DTs), Random Forest (RF) and Support Vector Machine (SVM) have been compared, with good results. Other analysis will be presented, like Principal Component Analysis (PCA) and clustering techniques, as a support to study student profiles.

II. Background Knowledge

This section will present and describe Affirmative Actions (*Ações Afirmativas*, as those brazilian specific legislations and projects for population support are called in portuguese. We

[1]Inep website: http://portal.inep.gov.br

[2]mean annual evasion: students who did not graduate, did not enroll in the following year (or the following semester, if the objective is to follow what happens in semester courses).

[3]Stanford Glossary: http://ai.stanford.edu/~ronnyk/glossary.html

will be concerned to substantiate the necessity for efficiency on these policies measures, and also describe other studies with our same goals.

A. Student Permanence, Affirmative Action and Evasion

The United Nations Educational, Scientific and Cultural Organization (UNESCO) *"Constructing an indicator system or scoreboard for higher education"* guide[4] presents at page 39 that the evasion rate is one of the main indicators of intern institution efficiency, alongside: graduation rate for a first qualification in higher education, and success rate for average years spent in higher education.

At Inep website related to OECD[5] we have access to different indicators, but related only to the aspect of costs and expenses. The lack of centralized data or evasion indicators is worrying, and government depends on independent auditors or usually, on a control made by the university itself.

The Brazilian government establish by laws the directives of university evaluation systems. The Lei de Diretrizes e Bases da Educação Nacional (LDB) (Law nº 9.394[6]), dated 12/20/1996, defines the guidelines and bases of national education. Specifically, the institutional evaluation is supported by Law no. 10.861 [7], dated 04/14/2004, which instituted Sistema Nacional de Avaliação da Educação Superior (SINAES). This project together with the Plano de Desenvolvimento Institucional (PDI) are the main regulatory processes of the higher education institutions by the State, and determine how higher education institutions are to be examined.

Specifically, the institutional evaluation is supported by Law no. 10.861 [8], dated 04/14/2004, which instituted the SINAES. This together with the PDI are the main regulatory processes of the higher education institution by the State, and provide how it should be examined [6]. 2016-2020 UNIFESP's PDI [9] reinforces (page 243) "the realization of the PDI will only be achieved if it involves permanent processes of institutional self-evaluation". Among the evaluations included in the document are actions such as (page 158) the preparation of socioeconomic studies, comparative student companionship (quota holders and non-quota holders), forums and debates involving the theme of "staying in the public university".

The Núcleo de Apoio ao Estudante (NAE) (student support department) is the front line to support UNIFESP students. It is bound to Pró-Reitoria de Assuntos Estudantis (PRAE) and seeks to apply the Política de Assistência Estudantil (PAE) defined by the Conselho de Assuntos Estudantis (CAE). And one of its first competences is the promotion of actions to contribute to the Políticas de Permanência Estudantil (PPE). The PAE does not exist as a single document, but it is a group of policies that coordinate a diverse set of actions for students support.

[4]UNESCO guide available at: http://bit.ly/2mnqb6W
[5]OECD Inep website: http://bit.ly/2lnEc4M.
[6]Law nº 9.394 available at: http://bit.ly/1SXIu8A
[7]Law no. 10.861 available at: http://bit.ly/2lFT5M6
[8]Law no. 10.861 available at: http://bit.ly/2lFT5M6
[9]PDI available at: http://bit.ly/2m20F4O

The second clause of Article 5 of the PRAE regiment [7], establishes the CAE that has the competence to formulate the Unifesp Student Support Policy.

The organizational structure of the PRAE has four coordinates: the Coordenadoria de Ações Afirmativas e Políticas de Permanência (Caap), the Coordenadoria de Atenção à Saúde do Estudante (Case), the Coordenadoria de Apoio Pedagógico e Atividades Complementares (Capac) and Coordenadoria de Cultura, Atividade Física e Lazer (Ccafl). These four coordinates have their responsibilities as defined in Article 11 in the PRAE regiment [7] and repeated at the Caap website[10], citing, in particular, his fifth assignment is: *raise data that contributes to the design of the socioeconomic and cultural profile of Community of Unifesp students, contributing to the continuous development of intra and inter institutional actions*.

One of the main problems is to standardize everything that concerns the Evasion [1]. It is a complex job to manage the flow of enrollments and obtain dropout reasons for each student. The Instituto de Ciência e Tecnologia (ICT) applies the Interdisciplinary Bachelor System in the early years, and since all students share the same initial course, specific indicators for certain courses would initially be impossible to acquire.

B. Machine Learning and Student Evasion

Alpaydin [8] displays Machine Learning as a step of **Data Mining**, where there is a quantity of data and you want to extract patterns. Machine Learning techniques can be divided by how they learn in three broad areas: supervised learning, unsupervised learning, and reinforcement learning. Machine learning techniques can also be divided by what type of problem it is solving. There are basically two types of problems, classification problems and regression problems. Supervised classification and non-supervised techniques are used in this project. There are several studies applying computational methods to the problem of university dropout [9]–[11] and regression of grades in crucial subjects [12].

The students *Fernando Dias and Peter Jandl* [9] present interesting statistical techniques for the analysis of risk of evasion in the *Information System* course of *Anchieta University*. To perform the analysis they used the technologies PHP, Apache, MySql and Google Charts with which they obtained some visualizations tools that would help teachers and coordinators of each course to follow the numbers related to the evasion of students.

Digiampietri et al. [12] analyzed 1,027 students, of which 627 (61%) graduates and 400 canceled enrollments (evasion) of students and alumni of the Bachelor of Information Systems EACH-USP in the period from 2005 to 2015. In order to carry out the analysis, they used the *Rotation Forest* classifier, with cross-validation in ten subsets, reaching a precision of 91.63% in predicting student scores in some specific disciplines.

Delen [10] has conducted studies to predict friction in students academic lives. Where they used a database of a public university in the Midwest, with a median enrollment of

[10]Caap website: http://bit.ly/2m0ZioD

23,000 students, and an average length of stay of the first year of 80%. The algorithms DT, SVM, Neural Networks (NN) and Logistic Regression (LR) were tested, and the best result was the DT algorithm with 87.23% accuracy on predicting .

Aulck [11] deals with a problematic dropout rate of 30% of first year undergraduates at American universities. The author used a balanced database with 32,538 students with the algorithms LR, RF and K-Nearest Neighbors (KNN), where LR reached the best precision of 66.59% and a Area under the curve (AUC) [11] of 0.729.

III. Materials and methods

Description of the experimental data set - The database includes 25 attributes (listed in the ANNEXES) of 7156 students enrolled in the years 2012, 2013 and 2014 on the campus of São Paulo, Diadema, Baixada Santista, São José dos Campos , Osasco and Guarulhos of UNIFESP. Each student has a column indicating whether or not the course has been abandoned, which will be the goal of the Machine Learning technique. An important detail is that there is only 1 attribute directly linked to the student academic life (*Grade Coefficient*), while the other 24 attributes are relative to personal characteristics. On this basis there are 5019 students who have completed or are still active in their courses, and 2137 students who have dropped out of their courses, resulting in a drop of 29.86% of students dropping out.

Data preprocessing - The main transformation in data, and most common, is to make non-numeric data into numeric data. For binary responses (eg yes or no) the substitution was made for 1 or 0 (zero), and for categorical responses the *1 to n* technique was used. New columns were generated with the possible values and were filled with 1s and 0s. For example, the column *Marital Status*, where the options are *married*, *separated*, *single* or *widower*, was distributed in four new columns *civil_state_single*, *civil_state_married*, and so on. After these transformations we obtained 75 columns in our database.

With the database being processed, the database was split into training and testing, with 5009 (approximately 75%) students for training and 2147 (approximately 25%) for testing, and a seed of randomness equal to 42.

Machine Learning Model Construction - For the code implementation of this project we used the Python programming language, a language developed under an open license - license approved by Open Source Initiative (OSI) - its use is free, even for commercial use. The license is managed by Python Software Foundation (PSF). The version used in this project was 3.5.2, and the main reason for choosing Python is the **scikit-learn** package[12] [13], a package with several tools for data mining and data analysis, also with free license - license Berkeley Software Distribution (BSD).

[11]common statistical measure in binary classification problems, contained in the range $[0, 1]$ where 1 represents the best possible result.

[12]official website of the scikit-learn package: http://scikit-learn.org/stable/

The following algorithms available in the **scikit-learn** package, with some details explained in the [13] documentation, were applied:

- **GNB**: is famous for good results in document classification and spam filtering. It can be extremely fast compared to other methods, and the dissociation of the distribution of the conditional characteristic of the class can be estimated independently as a one-dimensional distribution, alleviating the curse of dimensionality. Although it is a decent classifier, it is known to be a bad estimator, so probability outlets should not be taken too seriously.
- **DTs**: is a set of non-parametric supervised learning methods. The technique creates a model by learning simple decision rules inferred from the characteristics of the data. As DTs are simple to understand and interpret - can be viewed. They tend to learn too much (overfitting) and requires techniques to improve their generalization, and need balanced databases to avoid bias.
- **RFs**: a random forest is a meta estimator that fits a number of DTs on various sub-samples of the dataset and use averaging to improve accuracy and control overfitting. Each tree in the ensemble is built from a sample drawn with replacement from the training set. As a result of this randomness, the bias of the forest usually slightly increases (with respect to a single DT) but, due to averaging, its variance also decreases, usually compensating the bias increase, yielding a better model.
- **SVM**: started in OCR (optical character recognition) applications and later became reference in object recognition [14]. It is a set of supervised learning algorithms used for classification, regression and detection of outliers. The choice of SVM was due to its effectiveness in high dimensionality spaces. But SVMs do not directly provide estimates of probability, they are calculated using a 5-fold cross validation, which is computationally expensive.

In order to evaluate the models, time measurements and the statistics *Brier*, *F1 score*, *precision* and *recall* will be used. The Brier statistic is appropriate to evaluate binary and categorical results. Across all items in a set of predictions N, the Brier score measures the mean quadratic difference between (1) the predicted probability assigned to the possible outcomes for the Item i, and (2) the actual result. F1 score is a harmonic mean between precision p and, recall r, in the range $[0, 1]$ being 1 the best value, according to **Equation 1**, where tp *true positives*, fp *false positives* and fn *false negatives*. Precision is intuitively the ability of the model to not classify as positive an example that is negative and recall that the ability to find all positive examples.

$$\mathrm{F1} = 2 \cdot \frac{p \cdot r}{p + r} \tag{1}$$

$$p = \frac{tp}{tp + fp} \qquad r = \frac{tp}{tp + fn}$$

For analyzing clusters on the dataset it was applied two different types of non-supervised techniques:

- **PCA**: is used to decompose a multivariate dataset in a set of successive orthogonal components that explains a maximum amount of the variance. The dimensions after applying PCA to a dataset is a ordered list of new features (combinations of old features) that explains a percentage of the original dataset variance, ordered in a way that the first new features explains the most variance.
- **k-means clustering algorithm (Kmeans)**: this algorithm requires the number of clusters to be specified, and it will cluster data by trying to separate samples in n groups of equal variance. It will divide a set of samples by the mean of the samples in the clusters. This means are called "centroids", and they are not points from the original data, although they serve to represent a cluster.

From the techniques above we have enabled some possibilities, (1) reduce the 75 dimensional data to 2 dimensions, so we can visualize it, with caution of the original explained variance within this 2 dimensions, (2) analyze how many clusters are present in our original dataset and in our new PCA transformed dataset with this goal we can then (3) distinguish if the population in the study is split in representative clusters, how many cluster and what distinguish them.

For evaluating the Kmeans results we will apply the **Silhouette** statistic. It refers to the consistency within the clusters of data, or, how similar an object is to its own clusters (cohesion) compared to other clusters (separation). Silhouette ranges from $[-1, 1]$, where higher values indicate that the objects are well matched to its own clusters and poorly matched to neighboring clusters. It's defined, for one data sample i, by **Equation 2**

$$\text{s(i)} = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \tag{2}$$

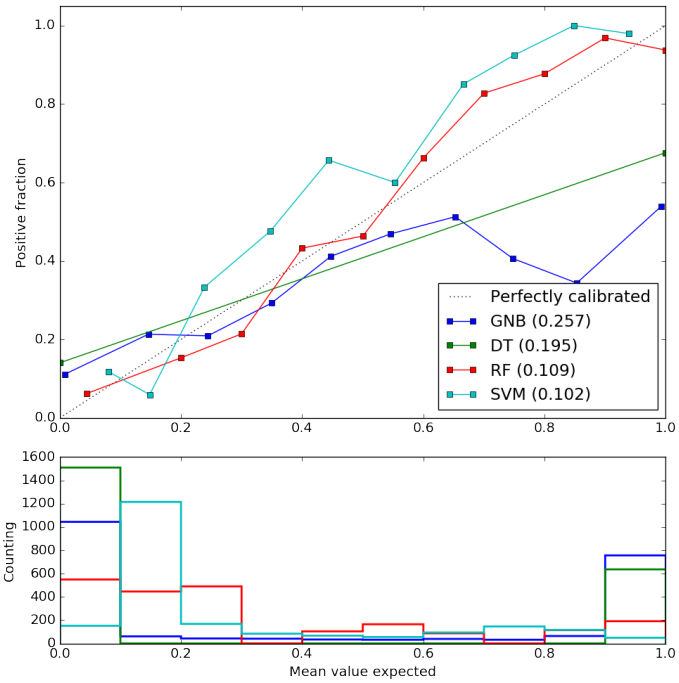
Where $b(i)$ is the lowest dissimilarity of i to any other cluster, of which i is not member, and $a(i)$ is the average dissimilarity of i with all other data within the same cluster. To obtain the final Silhouette, it is computed the average over all data samples.

The result of the application and analysis are explained in the next section.

IV. Results

The training database (with 5010 students) was separated into 3, with 1670, 3340 and 5010 individuals. The division was made in a random way and maintaining the same bases generated for the three algorithms. The dropout student was considered the positive class for calculating the statistics. **Table I** shows the obtained results, and analyzed in the next section.

Figure 1: Calibration graph (confidence curve)



A calibration process **Figure 1**, a way of accurately measuring a measurement system, has been performed in order to evaluate the degree of agreement between the result of a measurement and a conventional true value of the model. With this two experiments we will be able to decide which model, within the options, had the best efficiency and inserts the less bias for predicting student evasion.

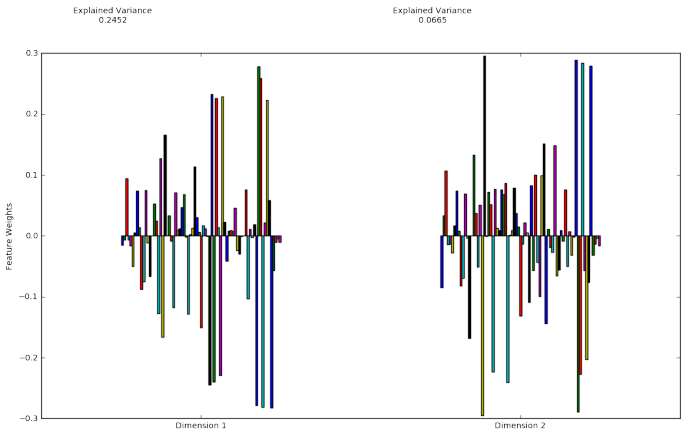
For the next experiment, analyzing student profile, first the the PCA was applied to the pre-processed 75 dimensional dataset, and turned into a 2 dimensional dataset. This two dimensions, as seen in **Figure 2** explains around 30% of the original variance.

Figure 2 also shows the composition of this 2 new dimen-

method	train set	training time	prediction time	F1 Score (training)	F1 Score (test)	*precision*	*recall*	Brier
GNB	1670	0.0054 sec	0.0022 sec	0.5648	0.5767			
	3340	0.0096 sec	0.0038 sec	0.6028	0.6260			
	5010	0.0113 sec	0.0035 sec	0.5919	0.6109	0.518	0.745	0.257
DT	1670	0.0152 sec	0.0008 sec	1.0000	0.6761			
	3340	0.0244 sec	0.0013 sec	1.0000	0.6761			
	5010	0.0361 sec	0.0035 sec	1.0000	0.6729	0.676	0.670	0.195
RF	1670	0.0250 sec	0.0045 sec	0.9867	0.7282			
	3340	0.0403 sec	0.0063 sec	0.9776	0.7169			
	5010	0.0583 sec	0.0127 sec	0.9785	0.7335	0.676	0.670	0.111
SVM	1670	0.1430 sec	0.0839 sec	0.8179	0.7788			
	3340	0.4725 sec	0.3181 sec	0.8068	0.7736			
	5010	0.8887 sec	0.6483 sec	0.8049	0.7660	0.876	0.681	0.102

Table I: Comparative table of results.

Figure 2: 2 PCA dimensions



sions. Each bar represents the original dataset feature weight over the new dimension 1 and 2, the half above is a positive weight, and the half below is a negative weight over the final vector representing Dimension 1 and 2. No more analysis were made over this independent weights, but might be useful for obtaining more details over student's profile.

For defining a number of clusters for the Kmeans we computed the *Silhouette* statistic iterating from 2 to 9 clusters found by Kmeans. The results are shown in **Table II**.

The highest values of Silhouette happens for 2 clusters, for the pca dataset (with 2 dimensions) and for the pre-processed dataset (with 75 dimensions). From this conclusion we computed the visualization shown in **Figure 3**, from which we can clearly see that there is two clusters of students. From this visualization there is not much more to accomplish, we now need to analyze each of the clusters to see in what they differ.

	Silhouette	
k	pca dataset	pre-processed dataset
2	0.5769	0.5141
3	0.5541	0.3136
4	0.5439	0.2761
5	0.5260	0.2768
6	0.5023	0.2251
7	0.4664	0.2235
8	0.4424	0.2226
9	0.4380	0.2049

Table II: Silhouette values for 2 to 9 clusters found by Kmeans for pca and pre-processed dataset

The number of students in each cluster is 3,252 and 3,904, and the main differences of the mode for each clusters are presented in **Table III** that will be discussed in the next section.

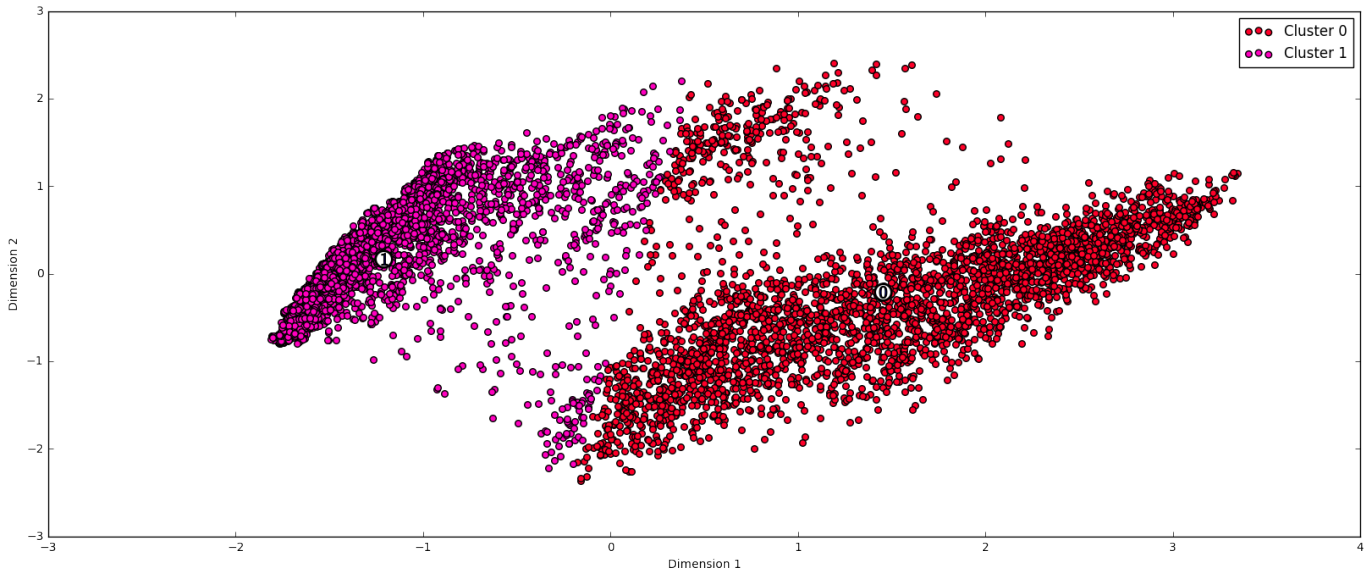
As the last statistic we analyzed the percentage of evasion for each of this clusters. And the percentage are, for cluster (0) (3,904 students) is 31.63% and for cluster (1) (3,252 students) is 37.98%

V. Discussion

The best F1 result obtained was the algorithm SVM with 0.7788. It obtained the best result, but its processing time, both training and prediction, are also the biggest among the chosen techniques. Although 15 times slower than the second slower technique (RF), it still below 1 second. The time of training and prediction are important when dealing with a large amount of data, it is noticed that GNB obtained the smallest processing time, but its F1 score is the least among the techniques. The two techniques with decision trees DT and RF (a ensemble of DTs) underwent overfitting[13]) in training and consequently obtained worse results in the test phase.

[13]when the statistical model fits in too much to the training dataset.

Figure 3: Cluster Learning on PCA-Reduced Data - Centroids Marked by Number



#cluster	Q5 private (1) or public (0) school	Q7 have (1) or do not have (0) a college degree	Q10 parents (1) or student (0) are/is responsible	Q11 working (1) or not working(0)	Q12 work without registration (1)	Q14 $>$ 1 hour to arrive at college	never worked (1)	income / minimum wage
3904	1	1	1	0	0	0	1	8.5
3252	0	0	0	1	1	1	0	3.5

Table III: Main differences between kmeans clusters modes.

For the clustering and analysis experiment we can clearly conclude that the two clusters found differ in meaningful ways, in view of quality of life and resources. In **Table III** we can see main differences between the $income/minimum\ wage$ feature, indicating a less financial resourceful group, question 14 also indicates that the less resourceful group need to travel for more than 1 hour to arrive at the University, probably harmful over their life quality. One crucial point also is question 12, that points out that at least half of the students in the less resourceful group works without proper registration, we as authors point out this grave problem and expect fast actions or programs from the University to fight this reality. The clusters also differ significantly over the dropout rate in each group, that might indicate a causality. But this means that NAE support and Affirmative actions can be helpful, nonetheless the support needs to reach those students and other supports dealing with life quality might be also needed.

VI. Conclusion

We think that the use of Machine Learning tools and analysis can help on strategic decisions for student support programs. This techniques can with great certainty identify a profile for a dropout student (F1 score 0,7788), and visualization techniques can drive specific actions that can be more efficient.

Acknowledgements

The authors thank the Institute of Science and Technology of the Federal University of São Paulo, where good people motivated this project.

References

[1] M. B. d. C. M. Lobo, "Panorama da evasão no ensino superior brasileiro: aspectos gerais das causas e soluções," *Associação Brasileira de Mantenedoras de Ensino Superior. Cadernos*, no. 25, 2012.

[2] C. APARECIDA, S. BAGGI, and D. A. LOPES, "Evasão e avaliação institucional no ensino superior: uma discussão bibliográfica," 2011.

[3] S. A. J. POLYDORO, "Evasão em uma instituição de ensino superior: desafios para a psicologia escolar." Ph.D. dissertation, Dissertação (Mestrado em Psicologia)–Departamento de Pós-Graduação em Psicologia da Pontifícia Universidade Católica de Campinas, Campinas, 1995.

[4] R. L. L. Silva Filho, P. R. Motejunas, O. Hipólito, and M. Lobo, "A evasão no ensino superior brasileiro," *Cadernos de pesquisa*, vol. 37, no. 132, pp. 641–659, 2007.

[5] R. A. Ambiel, "Construção da escala de motivos para evasão do ensino superior," *Avaliação Psicológica*, vol. 14, no. 1, pp. 41–52, 2015.

[6] G. Marback Neto, "Avaliação: instrumento de gestão universitária," *Vila Velha: Hoper*, 2007.

[7] "PRAE Regimento," http://bit.ly/2lFJQf3, acesso em: 18-09-2016.

[8] E. Alpaydin, *Introduction to machine learning*. MIT press, 2014.

[9] F. S. Dias and P. J. Junior, "Ferramenta para avaliação de risco de evasão acadêmica," *Revista Engenho*, vol. 12, 2016.

[10] D. Delen, "A comparative analysis of machine learning techniques for student retention management," *Decision Support Systems*, vol. 49, no. 4, pp. 498–506, 2010.

[11] L. Aulck, N. Velagapudi, J. Blumenstock, and J. West, "Predicting student dropout in higher education," *arXiv preprint arXiv:1606.06364*, 2016.

[12] L. A. Digiampietri, F. Nakano, and M. de Souza Lauretto, "Mineração de dados para identificação de alunos com alto risco de evasão: Um estudo de caso," *Revista de Graduação USP*, vol. 1, no. 1, pp. 17–23, 2016.

[13] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux, "API design for machine learning software: experiences from the scikit-learn project," in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 2013, pp. 108–122.

[14] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and computing*, vol. 14, no. 3, pp. 199–222, 2004.

GLOSSARY

AUC
: Area under the curve. 3

BSD
: Berkeley Software Distribution. 3

Caap
: Coordenadoria de Ações Afirmativas e Políticas de Permanência. 2

CAE
: Conselho de Assuntos Estudantis. 2

Capac
: Coordenadoria de Apoio Pedagógico e Atividades Complementares. 2

Case
: Coordenadoria de Atenção à Saúde do Estudante. 2

Ccafl
: Coordenadoria de Cultura, Atividade Física e Lazer. 2

DT
: Decision Tree. 1, 3–5

GNB
: Gaussian Naive Bayes. 1, 3–5

ICT
: Instituto de Ciência e Tecnologia. 2

Inep
: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. 1, 2

Kmeans
: k-means clustering algorithm. 4, 5

KNN
: K-Nearest Neighbors. 3

LDB
: Lei de Diretrizes e Bases da Educação Nacional. 2

LR
: Logistic Regression. 3

MEC
: Ministério da Educação. 1

NAE
: Núcleo de Apoio ao Estudante. 2, 6

NN
: Neural Networks. 3

OECD
: Organization for Economic Co-operation and Development. 1, 2

OSI
: Open Source Initiative. 3

PAE
: Política de Assistência Estudantil. 2

PCA
: Principal Component Analysis. 1, 4

PDI
: Plano de Desenvolvimento Institucional. 2

PPE
: Política de Permanência Estudantil. 2

PRAE
: Pró-Reitoria de Assuntos Estudantis. 2

PSF
: Python Software Foundation. 3

RF
: Random Forest. 1, 3–5

SINAES
: Sistema Nacional de Avaliação da Educação Superior. 2

SVM
: Support Vector Machine. 1, 3–5

UNESCO
: United Nations Educational, Scientific and Cultural Organization. 2

UNIFESP
: Universidade Federal de São Paulo. 1–3

ANNEXES

Original dataset features:

Q1 - Estado Civil (nominal: ”casado”, ”separado”, ”solteiro” ou ”viúvo”)
Q2 - Filhos (binário: ”Tem” ou ”Não tem”)
Q3 - Perfil Étnico-Racial (nominal: ”amarelo”, ”branco”, ”indígena”, ”pardo” ou ”preto”)
Q4 - Turno no Ensino Médio (binário: ”diurno” ou ”noturno”)
Q5 - Tipo de Ensino Médio (nominal: ”particular”, ”pública” ou ”exterior”)
Q6 - Escolaridade do Pai (nominal: ”fundamental incompleto”, ”fundamental completo/ensino médio incompleto”, ”médio completo/superior incompleto”, ”superior completo”, ”não sei” ou ”sem instrução/não alfabetizado”)
Q7 - Escolaridade da Mãe (nominal: ”fundamental incompleto”, ”fundamental completo/ensino médio incompleto”, ”médio completo/superior incompleto”, ”superior completo”, ”não sei” ou ”sem instrução/não alfabetizado”)
Q8 - Principal responsável pela renda (nominal: ”eu”, ”irmão/irmã”, ”cônjuge ou companheiro”, ”filho ou outra pessoa” ou ”meus pais”)
Q9 - Contribuintes da renda familiar (nominal: ”acima de quatro pessoas”, ”quatro pessoas”, ”três pessoas”, duas pessoas” ou ”uma pessoa”)
Q10 - Responsável pela manutenção na universidade (nominal: ”companheiro/cônjuge”, ”outra pessoa”, ”pais”, ”irmão/irmã”, ”eu”)
Q11 - Situação de trabalho do estudante (nominal: ”desempregado”, ”empregado”, ”empregador”, ”nunca trabalhou” ou ”outra situação”)
Q12 - Tipo de vínculo no trabalho (nominal: ”nunca trabalhei”, ”já trabalhei, mas não agora”, ”trabalho com ou sem carteira” o ”trabalho por conta própria”)
Q13 - Residência durante os estudos (nominal: ”com família na cidade do campus”, ”com família afastado mais de 60km do campus”, ”com família em cidade até 50km do campus”, ”com outros estudantes afastado mais de 20km do campus”, ”com outros estudantes em cidade até 15km do campus”, ”em moradia cedida ou própria dividindo despesas com outros estudantes” ou ”outro”)
Q14 - Tempo de deslocamento até a faculdade (nominal: ”no estado de São Paulo”, ”região sul”, ”região centro-oeste”, ”região nordeste”, ”região norte” ou ”em outro país”)
Q15 - Local de residência da família (nominal: ”no estado de São Paulo”, ”região sul”, ”região centro-oeste”, ”região nordeste”, ”região norte” ou ”em outro país”)
Q16 - Com quem mora atualmente (nominal: ”sozinho”, ”república, pensão, habitação coletiva, etc.”, ”com cônjuge ou companheiro”, ”com outros parentes”, ”com os pais” ou ”outra situação”)
Q17 - Atividade remunerada durante os estudos (nominal: ”não”, ”sim, de 1 a 2 anos”, ”sim, de 2 a 3 anos”, ”sim, mais de 3 anos”, ”sim, menos de 1 ano” ou ”sim, todo o tempo”
Q18 - Horas de trabalho durante os estudos (nominal: ”Nunca trabalhei”, ”Sem jornada fixa, até 10 horas semanais”, ”De 11 a 20 horas semanais”, ”De 21 a 30 horas semanais”, ”De 31 a 40 horas semanais” ou ”Mais de 40 horas semanais”)
Q19 - Idade com que começou a trabalhar (nominal: ”Nunca trabalhei enquanto estudava”, ”Antes dos 14 anos”, ”Após os 18 anos”, ”Entre 14 e 16 anos” ou ”Entre 17 e 18 anos”)
Q20 - Coeficiente de Rendimento (CR) (numérico no intervalo $[0, 10]$)
Q21 - Renda relativa ao salário mínimo (RSM) (numérico no intervalo $[0, \infty]$)
Q22 - Pessoas que vivem com essa renda (PVR) (numérico no intervalo $[0, \infty]$)
Q23 - Renda per capita (numérico: RSM/PVR)
Q24 - Campus (nominal: ”S.PAULO”, ”DIADEMA”, ”B.SANTISTA”, ”S.JOSÉ DOS CAMPOS”, ”OSASCO” ou ”GUARULHOS”)
Q25 - ANO DE ENTRADA (numérico: ”2012”, ”2014” ou ”2013”)