# PREVISÃO DE ABANDONO DE ALUNOS NUMA INSTITUIÇÃO DE ENSINO SUPERIOR

## *PREDICTING STUDENT DROPOUT IN A HIGHER EDUCATION INSTITUTION*

**Pedro Sobreiro[1]; Domingos Martinho[2]**
*[1]Instituto Politécnico de Santarém; [2]ISLA Santarém/I2ES*
sobreiro@esdrm.ipsantarem.pt ; domingos.martinho@islasantarem.pt

### Resumo

O abandono é um problema nas instituições de ensino superior que tem vindo a aumentar a sua visibilidade, onde se verifica pouca investigação recorrendo a técnicas de *Machine Learning.* Neste estudo procuramos desenvolver um modelo para prevermos o abandono dos alunos, utilizando os dados históricos dos alunos de uma instituição de ensino superior. Os dados foram utilizados nos algoritmos de classificação *Two class logistic regression*, *Two class boosted decision*, *Two class neural network* e *Two class support vector*, onde avaliarmos a sua exatidão através da matriz de confusão e análise do *Receiver Operating Characteristic* curve. Os resultados obtidos permitiram identificar *Two class neural network* como o mais adequado para os dados que estamos a tratar.
No entanto, verificamos que necessitamos de aumentar a representatividade do abandono na amostra e incorporarmos mais variáveis, como satisfação com a instituição e oportunidades de trabalho.

***Palavras-chave*:** abandono, ensino superior, machine learning.

### Abstract

The dropout is a problem in higher education institutions that has been increasing its visibility, area where is observed a lack of research using Machine Learning techniques. In this study, we try to develop a model to predict the dropout of students, using the historical data of a higher education institution students. The data were analyzed using the classification algorithms Two class logistic regression, Two class boosted decision, Two class neural network and Two class support vector, where we evaluate its accuracy using the confusion matrix and the analysis of the Receiver Operating Characteristic curve. The results obtained allowed to identify Two class neural network as the most suitable for the data that we are dealing with.
However, we observed the need to increase the representativity of dropout in the sample and incorporate more variables, such as satisfaction with the institution and job opportunities.

***Keywords*:** dropout, higher education, machine learning.

## 1. INTRODUÇÃO

O abandono de alunos no ensino superior é um problema que tem vindo a aumentar a sua visibilidade (Tinto, 2006). Portugal não é exceção, em março de 2013, o Governo português através de uma resolução da assembleia da República 60/2013, recomenda a elaboração de um relatório sobre o abandono nas instituições de ensino superior.

Apesar da maior visibilidade que o problema tem vindo a ganhar, verifica-se a existência de poucos estudos para prever o abandono de alunos (Bogard, Helbig, Huff, & James,

2011), onde o potencial na utilização de técnicas baseadas no *machine learning* não tem sido aproveitado (Aulck, Velagapudi, Blumenstock, & West, 2016). O abandono pode ser previsto com exatidão (Aulck et al., 2016), através da utilização de variáveis adequadas. As variáveis que explicam o abandono já foram identificadas (Bean, 1980; Ferreira & Fernandes, 2015).

O objetivo deste estudo é prever o abandono de alunos utilizando dados históricos de uma instituição de ensino superior.

## 2. METODOLOGIA

Os dados foram disponibilizados pelo responsável da instituição, contendo a informação de 280 alunos, com as variáveis: Idade; Género; Ano; Estado civil; Distância; Curso e Status. A distância foi calculada para cada aluno através de uma rotina que utiliza uma API para o google maps enviando o código postal da instituição e do aluno e tratando o resultado obtido em XML. No final procedemos à normalização dos dados numéricos (Idade; Ano e Distância).

O tratamento dos dados e as estatísticas descritivas foram realizadas num notebook em Jupyter (Ragan-Kelley et al., 2014) e a previsão através dos algoritmos de classificação foram desenvolvidos num modelo no Microsoft Azure Machine Learning Studio (ver Figura 4). Para construir a previsão foram utilizados 50% dos dados para treinar o modelo, sem o objetivo de realizar previsões. Os restantes 50% dos dados foram utilizados para testar o modelo, i.e. avaliar se a previsão permite antecipar com exatidão nos dados de teste, se a situação do aluno (status) corresponde à previsão realizada.

As previsões foram realizadas com algoritmos de classificação, para prever o resultado Ativo (aluno retido) ou Anulado (Aluno que abandonou o curso). Os algoritmos utilizados foram: *Two class logistic regression*; *Two class boosted decision*; *Two class neural network* e *Two class support vector*.

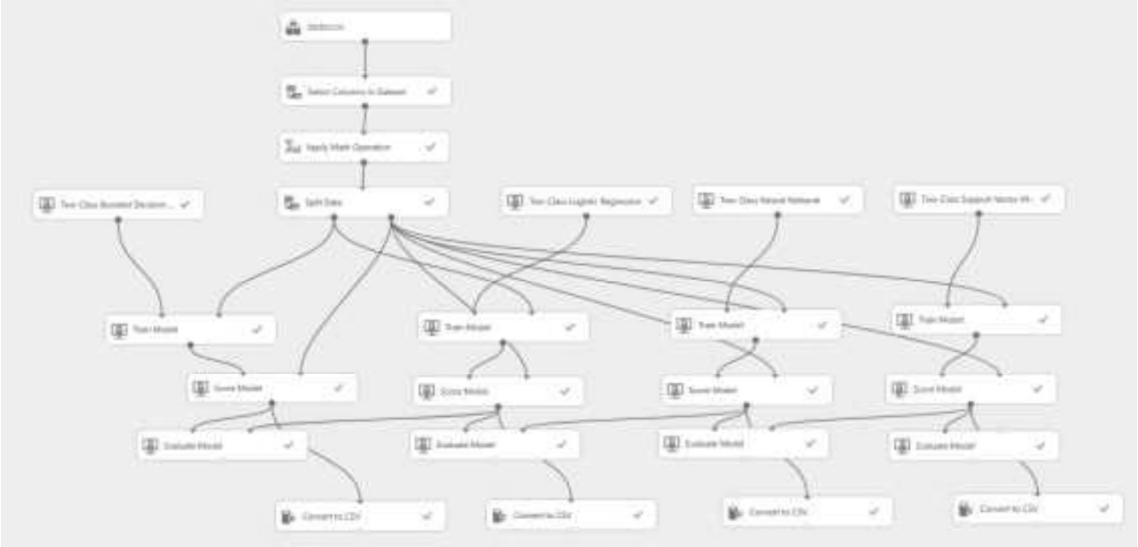

Figura 4. Modelo utilizado para correr os algoritmos

## 3. RESULTADOS

A caracterização das variáveis numéricas é apresentada no Quadro 1. A exatidão da previsão está representada no Quadro 2.

| | Idade | Ano | Distância |
|---|---|---|---|
| count | 280.000000 | 280.000000 | 280.000000 |
| mean | 30.271429 | 1.325000 | 44.817793 |
| std | 10.248415 | 0.620513 | 61.506057 |
| min | 19.000000 | 1.000000 | 0.000000 |
| 25% | 22.000000 | 1.000000 | 2.227750 |
| 50% | 27.000000 | 1.000000 | 21.709000 |
| 75% | 37.000000 | 1.000000 | 60.844750 |
| max | 69.000000 | 3.000000 | 309.868000 |

Quadro 1. Descrição das variáveis numéricas

A matriz de confusão (*Confusion Matrix*) representa os *True Positive* (Positivo com resultado previsto de positivo), *True Negative* (Negativo com resultado previsto de negativo), *False Positive* (Negativo com resultado previso de positivo) e *False Negative* (Positivo com resultado previsto negativo), estão representadas no Quadro 3. Os alunos ativos (Ativo) foram associados a um *label* Positivo e os que anularam a matricula (Anulado) a um *label* Negativo.

| MODELO | EXATIDÃO |
|---|---|
| **TWO CLASS LOGISTIC REGRESSION** | 0.971 |
| **TWO CLASS BOOSTED DECISION** | 0.936 |
| **TWO CLASS NEURAL NETWORK** | 0.971 |
| **TWO CLASS SUPPORT VECTOR** | 0.821 |

Quadro 2. Exatidão na previsão dos algoritmos de classificação utilizados

| | | | |
|---|---|---|---|
| True Positive | False Negative | True Positive | False Negative |
| **133** | **0** | **131** | **2** |
| False Positive | True Negative | False Positive | True Negative |
| **4** | **3** | **7** | **0** |
| **Two class neural network** | | Two class boosted decision | |
| True Positive | False Negative | True Positive | False Negative |
| **115** | **0** | **115** | **18** |
| False Positive | True Negative | False Positive | True Negative |
| **7** | **0** | **7** | **0** |
| **Two class logistic regression** | | Two class support vector machine | |

Quadro 3. Matriz de confusão obtida em cada um dos algoritmos utilizados

De acordo com os resultados obtidos, os algoritmos que apresentaram a maior exatidão foram *Two class logistic regression* e *Two class neural network*, com um valor de 0.971. Na Figura 5 podemos ver *Receiver Operating Characteristic* (ROC) curve dos algoritmos que apresentaram a maior exatidão, a linha vermelha representa *Two class neural network* e a azul *Two class logistic regression*. Quanto mais próxima a curva estiver do canto superior esquerdo, maior é a exatidão (Zweig & Campbell, 1993), o que significa que o algoritmo *Two class neural network* apresenta melhores resultados.

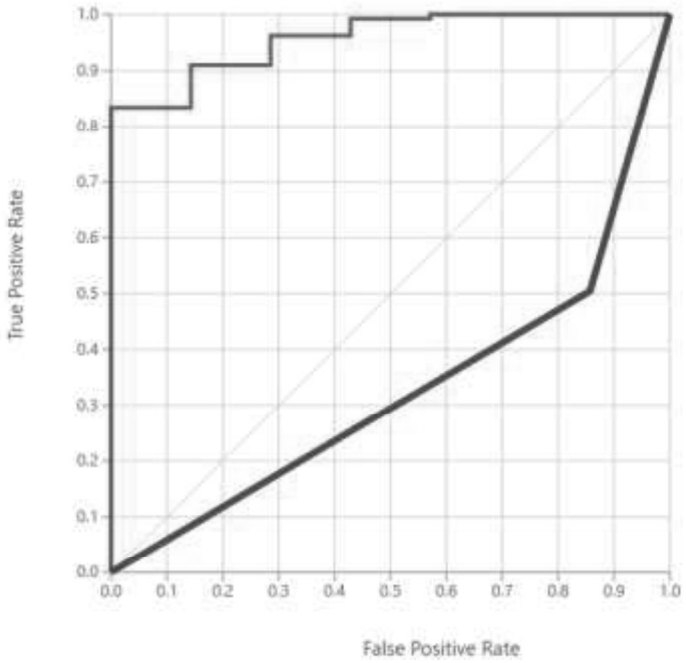


Figura 5. ROC curve Two class logistic regression (Azul) e Two class neural network

## 4. DISCUSSÃO DOS RESULTADOS

Os resultados obtidos permitiram identificar o algoritmo *Two class neural network* com uma exatidão muito elevada, no entanto podemos estar perante um problema em que o algoritmo pode erradamente criar relações irrelevantes entre os dados para criar a previsão (Moseley & Mead, 2008). A dimensão da amostra é fundamental, neste caso trabalhamos com uma amostra com um n=280, apesar de corresponder a valores aceitáveis (Figueroa, Zeng-Treitler, Kandula, & Ngo, 2012), o número de ocorrências de desistências na amostra é baixo (14 elementos), valor que pode ser insuficiente para treinar o modelo corretamente apesar da exatidão apresentada.

Bean (1980) sugere a utilização de variáveis como oportunidades de trabalho, satisfação e estatuto socioeconómico, variáveis que não foram utilizadas e que poderiam ser importantes para melhorar a previsão.

Conseguir prever corretamente algo passado, não garante que se consiga prever o futuro (Moseley & Mead, 2008). O algoritmo teria de ser testado em mais situações, antes de podermos afirmar que prevê corretamente o abandono ou retenção do aluno.

## 5. CONCLUSÃO

Neste estudo apresentamos resultados preliminares para prever o abandono dos estudantes de uma instituição do ensino superior. A amostra utilizada possuía uma dimensão de 280 elementos que participavam em vários cursos da instituição. Apesar

de obtermos uma exatidão elevada na previsão, o número de alunos que possuem um label de “Abandono” é baixo. O trabalho a desenvolver no futuro, passa por incorporar variáveis referentes às oportunidades de trabalho e satisfação com a instituição, e aumentar o número de elementos na amostra, de forma a existir uma maior representatividade no estudo do abandono.

## REFERÊNCIAS

Aulck, L., Velagapudi, N., Blumenstock, J., & West, J. (2016). Predicting Student Dropout in Higher Education. *arXiv preprint arXiv:1606.06364*. Obtido de https://arxiv.org/abs/1606.06364

Bean, J. P. (1980). Dropouts and turnover: The synthesis and test of a causal model of student attrition. *Research in higher education*, *12*(2), 155–187.

Bogard, M., Helbig, T., Huff, G., & James, C. (2011). A comparison of empirical models for predicting student retention. *White paper. Office of Institutional Research, Western Kentucky University*. Obtido de http://www4.wku.edu/instres/documents/comparison_of_empirical_models.pdf

Ferreira, F., & Fernandes, P. (2015). Fatores que influenciam o abandono no ensino superior e iniciativas para a sua prevenção: O olhar de estudantes. Obtido de http://www.fpce.up.pt/ciie/sites/default/files/ESC45Ferreira.pdf

Figueroa, R. L., Zeng-Treitler, Q., Kandula, S., & Ngo, L. H. (2012). Predicting sample size required for classification performance. *BMC Medical Informatics and Decision Making*, *12*, 8. https://doi.org/10.1186/1472-6947-12-8

Moseley, L. G., & Mead, D. M. (2008). Predicting who will drop out of nursing courses: a machine learning exercise. *Nurse Education Today*, *28*(4), 469–475. https://doi.org/10.1016/j.nedt.2007.07.012

Ragan-Kelley, M., Perez, F., Granger, B., Kluyver, T., Ivanov, P., Frederic, J., & Bussonier, M. (2014). The Jupyter/IPython architecture: a unified view of computational research, from interactive exploration to communication and publication. Em *AGU Fall Meeting Abstracts* (Vol. 1, p. 07). Obtido de http://adsabs.harvard.edu/abs/2014AGUFM.H44D..07R

Tinto, V. (2006). Research and practice of student retention: What next? *Journal of College Student Retention: Research, Theory & Practice*, *8*(1), 1–19.

Zweig, M. H., & Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, *39*(4), 561–577.