NOVA IMS Information Management School

MGI

Mestrado em Gestão de Informação
Master Program in Information Management

Using predictive modelling to create the school dropouts' profile

A case study regarding elementary and high school students

Maria Matilde de Magalhães Fechas Momade

Project Work report presented as partial requirement for obtaining the master's degree in Information Management

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

Using predictive modelling to create the school dropouts' profile

A case study regarding elementary and high school students

By

Maria Matilde de Magalhães Fechas Momade

Project Work report presented as partial requirement for obtaining the Master's degree in Information Management, with a specialization in Knowledge Management and Business Intelligence.

Advisor: Prof. Mauro Castelli, PHD

Fevereiro 2020

ABSTRACT

The students' disengagement with school is a worldwide contemporary topic, which has been to lengthy discussions. This event may be an indicator of the possibility of a precocious school dropout, becoming a burden for the students' families, schools and the Government.

This study focused only on a school located in Amadora, where the school dropout rate is quite significant. The main purpose of the present thesis is to understand students' proneness to quit school in a premature fashion. A dataset containing all the pupils' information available in that institution considered was transformed, trained and tested in order to produce a detailed analysis.

The main conclusions taken from the study are that the students' characteristics and familiar context play the major role in their likeliness to dropout school.

KEYWORDS

Education, Students, Parents, Community, School, Disengagement, Predictive Modelling, Machine Learning, Training Set, Test Set, Multicollinearity, Features, Logistic Regression, Random Forest, Performance, Sensibility, Specificity, AUC-ROC Curve

SUMÁRIO

O desinteresse escolar dos alunos com a escola, tópico de vastas discussões, é um tema atual em todo o mundo. Este fenónomo pode ser um indicador da possibilidade de abandono escolar precoce da escola, traduzindo-se num fardo para as famílias dos alunos, para as escolas e para o próprio Governo.

Este estudo focou-se somente numa escola localizada na Amadora, onde a taxa de abandono escolar é bastante significativa. O principal objetivo da presente tese é entender a propensão dos alunos para abandonar a abandonar a escola de maneira prematura. Um conjunto de dados que contém todas as informações dos alunos disponíveis na instituição considerada foi transformado, treinado e testado para produzir uma análise detalhada que procura responder à premissa base da investigação.

As principais conclusões tiradas do estudo são que as características e o contexto familiar dos alunos têm um papel determinante na sua probabilidade de abandonar a escola.

PALAVRAS-CHAVE

Educação, Estudantes, Pais, Comunidade, Escola, Desengajamento, Machine Learning, Modelos de Predição, Dados de Treinamento, Dados de Teste, Multicolinariedade, Recursos, Regressão logística, Random Forest, Desempenho, Sensibilidade, Especificidade, Curva ROC-AUC

TABLE OF CONTENTS

TABLE OF FIGURES

1. INTRODUCTION

1.1. Background and Problem Identification

In Modern society, educating can be both a time-consuming task and an increasing challenge when the school and parents' roles are confounded, arising different thoughts about whose responsibility it is to educate children.

School failure is undoubtedly an issue of great relevance in today's world, worsened by the implications it has in the professional and social life of people and communities. Consequently, the school performance of children and adolescents deserves greater attention and research in multiple areas such as psychology, sociology and educational sciences, being failure and dropout crucial issues both nationally and internationally.

Educational achievement and its relationship with the socioeconomic background is one of the most permanent issues in research. In Portugal, the socioeconomic conditions and familiar context play a significant role in the students' performance at school (Lobo, 2016). A myriad of national and international studies have shown that, on average, students who live under favoured socioeconomic backgrounds tend to achieve better school results than their peers who live under more disadvantageous backgrounds.

A study named "Desigualdades socioeconómicas e resultados escolares" conducted by Direcção-Geral de Estatísticas da Educação e Ciência has concluded that the Portuguese students' success at school is strongly related both to the socioeconomic level of their households and the academic qualifications of their mothers (Direcção-Geral de Estatísticas da Educação e Ciência, 2016; Casanova, 2016). It also points out that unfavourable socioeconomic indicators do not determine students' school failure.

According to INE – National Institute of Statistics –, despite the Portuguese early school dropout rate is above the European Union's (10.6%), it has decreased to its lowest, reaching 12.6% in 2017. Thus, the EU's target of a 10% rate in 2020 is closer than it was back in 2014, that is 17.4%. Still, the Ministry of Education states the reduction of the dropout rate should be kept as one of the main objectives of public policies. Considering this decrease, the national Minister of the Education, Tiago Brandão Rodrigues, congratulated schools and communities, professors and schools' employees for their contribution to make students carry out their compulsory schooling. Although Rodrigues attributes this to the school's mandatory attendance rate increase, he claims

the urge to implement more individualized measures to eradicate this phenomenon. In addition, the Minister reinforced that in Portugal female students are less likely to leave school early (9.7%) when compared to male ones (15.3%) (Marques Costa and Lusa, 2018).

1.2. Study Objectives

The objective of the present study is to measure, from a set of given variables, the above-mentioned disparities – that is, the main reasons behind the students' success at school – and in which sense these do relate with their decision to drop out school. Thus, conclusions will be taken concerning the profile of students who abandon high school.

The analysis focuses on students who attend the public and regular education, namely students enrolled in the third cycle of general basic education in Escola Básica 2/3 Dr. Azevedo Neves – a primary, elementary, preparatory and high school that has both regular and technical education (which include kitchen, pastry and catering; multimedia, design and fashion; and geriatrics). This school is located in Damaia, Amadora, one of the country's most densely populated county, and was considered as "the most African school in Europe" by Jornal de Notícias (Jornal de Notícias, 2018). The school is mainly frequented by students who belong to disadvantaged families from African countries where Portuguese is the official language and countries like Brazil, China, Romania, Ukraine, among others.

1.3. Study Relevance and Importance

The Direcção-Geral de Estatísticas da Educação e Ciência's study highlights that other external factors are affecting the students' success at school – besides the socioeconomic context and the mother's level of education (Direcção-Geral de Estatísticas da Educação e Ciência, 2016). It is then interesting to further explore those aspects, which counter the cause-effect relationship between the socioeconomic conditions of students' households and their success at school.

The collaboration and the responsibility of the community is key to building school success and commitment to education and valuing of learning. Therefore, the identification of the different specific aspects that contribute to teens dropping out of

school may provide valuable information for the School to act. Hence, the insights resulting from the analysis conducted can provide guidelines that support the effective decisions related to the School's efforts for increasing the retention levels of high school students, thereby reducing dropout rates.

2. LITERATURE REVIEW

2.1. Factors that influence students to drop out

According to the European Union's definition, early school leaving includes young students who have dropped out of school before the end of compulsory education but also those who have completed compulsory schooling but have not gained an upper secondary qualification.

Various are the theories conducted revolving the early dropout school rate. The main dissimilarities concerning the existing models regarding this topic are the importance given to the academic versus social behaviour and the gradual process of students' disengagement with school, finally driving them to drop out school.

Plenty of researches suggest that school dropout rates differ meaningfully according to one's both racial and ethnic background and socioeconomic context (Naylor, 1989; Lunenburg, 2000; National Center for Education Statistics, 2001).

This contemporary problematic has been given huge attention due to the vast number of students – belonging to minority groups – who leave school in advance (Rumberger, 1987).

In this chapter, this educational issue will be discussed at four different scopes, namely the student, family, community and school indicators.

2.1.1. Student-related indicators

In this topic, the individual perspective is the focus. It considers the *students' engagement*, including their values and behaviours and deduces to which extend these dictate their decision to drop out. Their behaviour includes both formal – such as classroom activities –, and informal aspects – such as the relationship with their peers and professors (Rumberger, 2001).

Recent theories converge by describing this issue as the final stage of the dynamic and cumulative process of school disengagement (Newmann et al., 1992; Wehlage et al., 1989) or withdrawal (Finn, 1989) from school. They also agree that there are two dimensions regarding the students' engagement with their educational establishment, which do weigh on their decision to leave school: the academic engagement and the social engagement. For instance, to stop doing homework is as valued as to not get along

with school peers.

Besides, according to the existing theories concerning school withdrawal, there are three major dimensions: academic achievements – which pertain to grades and evaluation scores –, educational stability – that indicate whether students remain in the same school or in school at all – and educational attainment – that is, the number of schooling years completed.

Academic achievement is a key dimension that affects the student's decision to school withdrawal. Having mediocre educational outcomes at school represents a major predictor of students dropping out (Ekstrom et al., 1986; Goldschmidt & Wang, 1999; Rumberger, 1995; Rumberger & Larson, 1998; Swanson & Schneider, 1999; Wehlage & Rutter, 1986). Several studies concluded that the early students' mean grades are a good predictor to deduct whether the student is going to withdrawal or not (Alexander et al., 1997; Barrington & Hendricks, 1989; Cairns et al., 1989; Ensminger & Slusacick, 1992; Garnier, Stein, & Jacobs, 1997; Morris, Ehren, & Lenz, 1991). In fact, the mean grades obtained in the fourth year of primary school are not only a good indicator to forecast the graduates' academic performance, but also to predict in which school year students are going to drop out (Roderick, 1993). The grades deterioration is, hence, an event occurring prior to leaving school; besides, the faster it occurs, the earlier the withdrawal will take place.

When it comes to students' stability, plenty of researchers posit it as both a cause and a consequence of students' engagement in school. Research suggest that students' mobility is a less severe form of student disengagement and dropout from school (Lee & Burkam, 1992; Rumberger & Larson, 1998) because it is believed that the risk of withdrawal is hugely influenced by students' school and residence mobility (Astone & McLanahan, 1994; Haveman et al., 1991; Rumberger, 1995; Rumberger & Larson, 1998; Swanson & Schneider, 1999; Teachman et al., 1996). To substantiate this theory, one study concerning high school inferred that, while the majority of graduates did not change school, the majority of students who did drop out had changed school at least once before the withdrawing (Rumberger et al., 1998).

Regarding educational attainment, absenteeism – the major indicator regarding the overall student's engagement (Bachman et al., 1971; Carbonaro, 1998; Ekstrom et al.,

1986; Goldschmidt & Wang, 1999; Rumberger, 1995; Rumberger & Larson, 1998; Swanson & Schneider, 1999; Wehlage & Rutter, 1986) – and retention are educational phenomena to pay attention to within the withdrawal scope. Though some studies posit retention may produce positive results on students' academic achievements (Alexander et al., 1994; Roderick et al., 1999), the majority of empirical researches declare that this event enhances the students' propension to withdrawal, regardless of the grade in which the retention took place (Goldschmidt & Wang, 1999; Grisson & Shepard, 1989; Jimerson, 1999; Kaufman & Bradby, 1992; Roderick, 1994; Roderick, Nagaoka, Bacon, & Easton, 2000; Rumberger, 1995; Rumberger & Larson, 1998). In addition, a 1995 study concluded that students retained in grades 1 to 8 were four times more likely to drop out between grades 8 to 10 than students who were not (Rumberger, 1995).

Finn (1989) had two perspectives on this topic. The first perspective, labelled "Frustrationself-esteem", suggests that an early school failure leads to the students' low self-esteem, resulting in the adoption of non-desirable behaviour. Hence, the student decreases his/her performance at school and so the cycle goes on until he/she decides to leave school, or is removed from it due to his/her bad behaviour. The second perspective, designated "Participation-Identification" model, defends both emotional and behavioural sphere commit to the withdrawal process. It advocates a poor involvement in school activities – such as homework, participation in class, responding to the teachers' directions and non-academic school activities – contributes to the students' low academic performance and their consequent misidentification with the institution. The major similarity between the models presented resides in the fact that both assume that dropping out school is a long-term process.

2.1.2. Family-related indicators

The socioeconomic status and familiar context play a major role in students' school achievements (Levin and Belfield, 2002). However, the environment in single parent houses (one "parental resource") – more than one-third of Hispanic households and more than two-thirds of African American households (KewalRamani et al., 2007) – is completely different when compared to the conceptual family structure.

The most obvious point of greater divergence consists of a reduced family income, once there is only one provider. A consequence of that is the fact that these families are

usually associated to a worse health status – which affects the children's capability to learn –, thanks to their poor access to both pre-natal care and health insurance, combined with their underprivileged nutrition (Wilder et al., 2008).

A family composed by two working parents may help children during their learning process, once they are more able to invest there, providing them with better conditions and access to better quality schools, with the ability to sign in extracurricular activities and summer schools. However, it may get to a point which an incremental positive impact on the income does not represent extra-educational benefits (Heckman, 2008). It is then possible to conclude that the students' performance at school does not have a linear relation with one's household financial situation.

Also, single-parent households and step families are more likely to drop out of school rather than students from the conceptual family structures (Astone & McLanahan, 1991; Ekstrom et al., 1986; Goldschmidt & Wang, 1999; McNeal, 1999; Rumberger, 1983; Rumberger, 1995; Rumberger & Larson, 1998; Teachman et al., 1996). On the other hand, concluded that a family that faces a couple separation is not increasingly likely to drop out school (Pong & Ju, 2000). Regarding the home environment itself, these households tend to not be a 'school-like' home – which means routines are not imposed, as in school, nor there are the materials needed to learn – and are more prone to have conflicts.

The parent-school interaction – which is a key player in the student's educational achievements – regarding students who dropout is low, because the first party is not likely to control its children attitudes towards the educational system nor their performance. For example, they are not likely to confirm if their kids did or did not do their homework.

In contrast, children whose parents have a strong relationship with school (McNeal, 1999; Teachman et al., 1996) and whose parents support and monitor school activities, encouraging them to make their own decisions – that is, authoritative parenting style – are less likely to dropout (Astone & McLanahan, 1991; Rumberger et al., 1990; Rumberger, 1995). Thus, parental involvement is crucial to a successful school route (Epstein, 1990; Suichu & Willms, 1996), once parents contribute to a decreasing likeliness to students' school withdrawal.

The human capital theory (Haveman & Wolfe, 1994) confirms the previously

mentioned theory by arguing that the children's educational ambition and cognitive skills depend on how their parents manage their time and other resources, such as money. Thus, the human and financial capital, parental education and parental income, respectively, explain the inherent connection between family background and school performance (Coleman, 1988).

2.1.3. Community-related indicators

Low-income children tend to live in poor areas, surrounded by a neighbourhood in which the criminality and the violence rates are higher. They are also more prone to have dropout friends, a fact that increases the likelihood of withdrawing too (Carbonaro, 1998).

The social capital within the community in which the children lives influences the quality of the school's resources and its learning system and, as an effect, to a low social capital community there are necessarily associated schools with poor learning resources (Brooks-Gunn et al., 1997; Hallinan & Williams, 1990; Wilson, 1987). This impacts the school's conditions in terms of danger and its learning resources' efficiency, such as the facilities themselves, the teaching quality and the availability of both mentoring and counselling systems.

Children who reside in low social capital resources communities are less likely to experience out-of-school educative experiences or pre-school, summer camps and extracurricular activities. This, adding to the fact that these areas lack cultural sources – such as museums and libraries –, leads to children preferring to watch TV rather than to read books or experience edifying activities.

Communities can instigate students to leave school by offering employment opportunities during or after school time. Even though it depends on the nature of the job and on the student's gender (McNeal, 1997a), working for long shifts while studying enhances the dropout odds rates (Bickel & Papagiannis, 1988; Clark, 1992, Rumberger, 1983).

2.1.4. School-related indicators

Wehlage believed two complex aspects are synergistically responsible for every school outcomes, including the dropout event – the *social bonding* and the *educational*

engagement. On the one hand, the *social bonding*, or school membership, concerns the involvement and commitment to the institution and its beliefs alongside one's social ties with its members. On the other hand, the *educational engagement* pertains to the academic dimension itself, which is affected by both the extrinsic rewards related to schoolwork and the intrinsic rewards related to the curriculum and how educational activities are founded (Wehlage et al., 1989).

Likewise, there are four important dimensions to consider regarding the school-related indicators. Three of them are quoted as *inputs* – resources, student composition, structural characteristic – and cannot be modified by the school itself (Hanushek, 1989), but only through policies; the fourth dimension represents the processes and practices. The latter can be monitored by the school itself and it is a tool to measure the school's effectiveness (Shavelson et al., 1987). This topic's conclusions are taken after controlling the different students' background characteristics.

Several studies found that, even after controlling both individual and contextual factors with possible impact on dropout rates, the student/teacher ratio had a positive and significant effect on both middle school and high school dropout rates (McNeal, 1997b; Rumberger, 1995; Rumberger & Thomas, 2000).

The school resources do influence the dropout rates. In fact, the higher the teachers' body quality is perceived by students, the lower the dropout rate; on the other hand, the higher the teachers' body quality is perceived by the school's principal, the higher the dropout rate (Rumberger & Thomas, 2000).

The students' body influences not only their performance as individuals but also influence to a collective extent (Gamoran, 1992), which makes this a predictor in the withdrawal decision (Bryk & Thum, 1989; McNeal, 1997b; Rumberger, 1995; Rumberger & Thomas, 2000).

Regarding the school structure dimension, major attention has been given to its structural feature (Bryk et al., 1993; Chubb & Moe, 1990; Coleman & Hoffer, 1987) – that is, public or private school – rather than to its size or location. In fact, studies have shown that public schools present higher dropout rates than catholic and other private schools (Bryk & Thum, 1989; Coleman & Hoffer, 1987; Evans & Schwab, 1995; Neal, 1997; Rumberger & Thomas, 2000; Sander & Krautman, 1995). Regarding the school's

dimensions, the smaller the school, the higher the likeliness to foster students' and staff's engagement (Wehlage, Rutter, Smith, Lesko, and Fernandez, 1989).

Lastly, school practices and policies together represent the most important dimension, being the one that can easily be acted upon. Both school's academic and social climate are predictors to the withdrawal rates (Bryk & Thum, 1989; Rumberger, 1995; Rumberger & Thomas, 2000), which can be measured, for example, by the attendance rates in classes. Educational institutions might increase the pupils' turnover by voluntarily or involuntarily instigating their withdrawal from school. The first occurs by the existence of general policies that affect students' engagement (Finn, 1989; Wehlage, Rutter, Smith, Lesko, and Fernandez, 1989) or by endangering favourable conditions for their engagement. The second occurs due to systematic exclusion, suspensions or forced transfers of problematic students (Bowditch, 1993; Fine, 1991; Riehl, 1999).

2.2. Costs and consequences of early school drop out

Early school dropout represents a loss of potential, that affects one's social and economic life's sphere. It is highly associated to reduced learnings, low social status and mental health deficit (Schoon, Duckworth, 2010).

Dropouts represent an additional charge to the Government that impacts the society in terms of both tax revenue loss and productivity (Naylor, 1989), resulting in higher costs of public services (Belfield & Levin, 2007).

Children abandoning school are more prone to have health problems and to engage in criminal activities (Rumberger, 1987).

3. METHODOLOGY

3.1. Data Treatment

3.1.1. Data Cleaning

Considering the previously defined sample – that is, students attending 9th to 12th grades –, the first step was to disregard all the data regarding students whose school grade is prior to the 9th.

In addition, some variables were immediately inconsiderate from the dataset for two main reasons: on the one hand, some students' sensible data was provided and should not be considered for data protection safeguard and, on the other hand, some variables do have a high correlation between them – multicollinearity (Frisch, 1994). Multicollinearity concerns independent variables that are so correlated that one can be used to predict the other one and vice-versa. This issue – the near-linear dependence – is seen as a violation of one of the basic assumptions for a successful regression model (Jason W. Osborne and Elaine Waters, 2002), once it might lead to a skewed analysis and hence, to a distorted model.

Consequently, the variables *Ano Lectivo, Processo, Num. Arquivo, Tipo Ident., Validade, SIGO, Swift/Bic, Freguesia de Naturalidade, Concelho de Naturalidade, Distrito de Naturalidade, Freguesia de Residência, Concelho de Residência, Distrito de Residência, Idade 15 Set, Data Sit., Escola,Turma, Ano na turma, Nível na turma, Data Matrícula, Nº Ordem, Ensino Articulado,Delegado, Subdelegado, AEC, Cod. Postal 4 E. Edu.,Cod. Postal 3 E. Edu.,Rep. dos Pais, Entidade Empregadora do Pai, Entidade Empregadora da Mãe,Data Última Vacina, Data Próxima Vacina, Data Carta Curso, Ano Letivo Conc. Curso, Média Curso* were excluded. The variable that nominates the students' gender – *Sexo* – was excluded so that conclusions would not be taken considering that feature.

The dependent variable – *Situação* – referring to the students' situation in school was changed from a categorical one to a binary one. Previously, it considered the following categories: 'X', meaning that a student is in a regular situation; 'TR' meaning that a student got transferred to another school; 'EF', meaning that he or she was excluded by absence; 'MT', meaning that the pupil was shifted class; 'AM'/'AS', meaning that the student cancelled his or her registration in school. Thus,

students who both had regular school situations, who changed classes and who were transferred to another school were considered to not be dropouts, unlike the remaining ones.

```
In [2]: print(dataset['Situação'].value_counts())
X     443
TR     52
EF     37
MT     23
AM     20
AS      1
Name: Situação, dtype: int64
```

Figure 1 – Series of unique values of the variable Situação

Likewise, the variables that distinguish students who do have a computer and do have access to the internet were converted into binary ones as well.

The education type was also converted into a binary variable to ease the models' appliance, in the sense that the Regular and the Professional education could be split. The first category is composed of students who are enrolled in the typical classes, being the second one composed of students who are enrolled in the school's professional courses.

```
In [3]: print(dataset['Tipo de Ensino'].value_counts())
Ensino Profissional                   307
Ensino Secundário                     102
3º Ciclo do Ensino Básico              77
Curso de Educação e Formação Tipo 2    61
Curso de Educação e Formação Tipo 3    29
Name: Tipo de Ensino, dtype: int64
```

Figure 2 – Series of unique values of the variable Tipo de

The designation of the students' school year is defined according to their education type. For example, if a student is in the professional education equivalent to the 10th grade, it is defined as '1 (10)' in the dataset. Therefore, the academic years were standardised, regardless of the students' education type.

```
In [10]: print(dataset['Ano'].value_counts())
1(10)        127
2(11)         93
3(12)         87
9º ano        77
CEF 2 - 1     61
10º ano       59
CEF 3         29
11º ano       29
12º ano       14
Name: Ano, dtype: int64
```

Figure 3 – Series of unique values of the variable Ano

The issue concerning the fact that multiple observations referred to the same course name was also solved:

```
In [11]: print(dataset['Nome Curso'].value_counts())
3º Ciclo                                                          77
Técnico de Cozinha/Pastelaria (A Partir De 2015/2016)             64
Curso Científico-Humanístico de Ciências e Tecnologias            58
Técnico de Restaurante/Bar (A Partir De 2015/2016)                47
Curso Científico-Humanístico de Línguas e Humanidades             44
Técnico de Massagem de Estética e Bem-Estar (A Partir De 2015/2016)  34
Técnico de Geriatria (A Partir De 2015/2016)                      33
Empregado/a de Restaurante/Bar - Empregado/a de Restaurante/Bar   32
Técnico de Multimédia (A Partir De 2015/2016)                     30
Operador/a de Informática - Operador/a de Informática             29
Agente em Geriatria - Trabalho Social e Orientação                29
Técnico de Restauração - Cozinha - Pastelaria (A Partir De 2006/2007)  24
Técnico de Design de Moda (A Partir De 2015/2016)                 21
Técnico de Gestão de Equipamentos Informáticos (A Partir De 2005/2006)  16
Técnico de Multimédia (A Partir De 2006/2007)                     14
Técnico de Design de Moda (A Partir De 2006/2007)                 13
Técnico de Restauração - Restaurante - Bar (A Partir De 2006/2007)  11
Name: Nome Curso, dtype: int64
```

Figure 4 – Series of unique values of the variable Nome Curso

In the raw dataset, there are plenty variables concerning the student's residence location. The downside is that they are all textual rather than numeric, except the one referring the proximity of the schoolhouse route – which is a great one to evaluate to which extent does the nearness of the school contributes to students' disengagement to the school.

```
In [16]: print(dataset['Tempo de percurso'].value_counts())
10    130
15    100
0      96
30     62
20     62
5      42
60     21
40     21
45     16
50      9
25      7
12      3
35      2
90      2
80      1
4       1
3       1
Name: Tempo de percurso, dtype: int64
```

Figure 5 – Series of unique values of the variable Tempo de percurso

This variable ranges from 0 to 90 minutes, being divided into three different classes: Fast (0 to 10 minutes), Medium (11 to 25 minutes), Far (26 to 45 minutes), Far Off (longer than 45 minutes).

The dataset indicates the students that might be eligible for three different types of social support, each with different echelons. These had dissimilar classifications:

```
In [21]: print(dataset['Escalão ASE'].value_counts())
<S/D>    188
A        154
B         69
C         10
--         7
Name: Escalão ASE, dtype: int64

In [22]: print(dataset['Escalão Abono'].value_counts())
--    341
1     157
2      68
3      10
Name: Escalão Abono, dtype: int64

In [23]: print(dataset['Escalão e-Iniciativas'].value_counts())
<S/D>    251
A         34
B         16
Name: Escalão e-Iniciativas, dtype: int64
```

Figure 6 – Series of unique values of the variables regarding social supports

Therefore, measures were taken to classify them the same way: 1 is equivalent to A; 2 is equivalent do B, C is equivalent to 3 and null means the student is not eligible for that support.

In addition, analysing the dataset using Python's properties makes it possible to conclude that there is a significant number of students from African countries – as expected, as this is “the most African school of Europe” –, and a lower percentage of students from other European and even Asiatic countries.

```
In [17]: print(dataset['Naturalidade'].value_counts())
Portugal                                          342
Cabo Verde                                        117
Guiné-Bissau                                       34
São Tomé e Príncipe                                25
Brasil                                             23
Angola                                             16
Reino Unido da Grã-Bretanha e Irlanda do Norte      4
Roménia                                             3
Nepal                                               2
Espanha                                             2
Estados Unidos da América                           2
Bielo-Rússia                                        1
Nigéria                                             1
França                                              1
Canadá                                              1
China                                               1
Ucrânia                                             1
Name: Naturalidade, dtype: int64
```

Figure 7 – Series of unique values of the variable Naturalidade

Both students, parents and caregivers' naturalness and nationality were classified according to their respective Continent, except the European one that was split into 'Portugal' and 'Rest of Europe'.

The academic background of students' parents and caregivers' was divided into four categories – High, Medium, Low and Other/Unknown –, according to their level of studies. Parents and caregivers who have a bachelor, baccalaureate, postgraduate or master's degree are included in the first category; those who completed high school are considered to have a medium level of education; those who have no literary abilities, or have no education higher than the 9th grade are considered to have a low level of education. The remaining cases were categorized as 'Other/Unknown'.

The parents' professional situation was classified into two different groups. The first one regards those who are self-employed, employed or isolated workers; and remaining one regards unemployed, retired, domestic and others.

```
In [14]: print(dataset['Sit. Emprego do Pai'].value_counts())
Situação Desconhecida                          296
Trabalhador por conta de outrem                168
Desempregado                                    74
Trabalhador por conta própria como isolado      10
Reformado                                        8
Trabalhador por conta própria como empregador    7
Outra                                            5
Name: Sit. Emprego do Pai, dtype: int64

In [15]: print(dataset['Sit. Emprego da Mãe'].value_counts())
Trabalhador por conta de outrem                251
Situação Desconhecida                          189
Desempregado                                    69
Doméstico                                       32
Trabalhador por conta própria como isolado      19
Outra                                            5
Trabalhador por conta própria como empregador    4
Reformado                                        4
Estudante                                        1
Name: Sit. Emprego da Mãe, dtype: int64
```

Figure 8 – Series of unique values of the variables Sit. Emprego do Pai e Sit.Emprego da Mãe

There are also other variables regarding the parents' occupation and professional area but these, especially the first one, have multiple missing values and do not add much to the previously stated variables; so, the highlighted variables were excluded from the dataset.

```
Profissão do Pai                1.215278
Area Profissional do Pai       53.472222
Sit. Emprego do Pai             1.388889
Form. Académica do Pai          0.000000
Cod. Postal 4 da Mãe           18.055556
Cod. Postal 3 da Mãe           18.055556
Cod. Postal Localidade da Mãe  17.361111
Nacionalidade da Mãe            0.173611
Naturalidade da Mãe             0.173611
Profissão da Mãe                0.173611
Area Profissional da Mãe       39.062500
```

Figure 9 – List of irrelevant variables with missing values

Still, within the caregivers' scope, some students are to their parents' responsibility, to themselves, to their grandparents, to their aunts or uncles, to their siblings or even to their tutors. Thus, the caregivers' parenting level was split into three categories: the parents, the student him/herself and others.

```
In [9]: print(dataset['Parentesco do E. Edu.'].value_counts())
Mãe        380
Pai         85
Próprio     51
Outro       19
Avó         13
Tia          9
Irmã         4
Tutor        4
Tio          3
Avô          2
Irmão        2
Name: Parentesco do E. Edu., dtype: int64
```

Figure 10 – Series of unique values of the variable Parentesco do E. Edu.

3.1.2. Dealing with missing values

Data's absence might bring problems to the data analysis method; hence, the missing values' issue must be solved prior to the models' application. In effect, careful data analysis is demanded to avoid this issue. Additionally, in order to solve this, it is important to study the variable's distribution.

Using Python's properties, missing values were found in the variables presented below:

```
Sit. Emprego do Pai          1.388889
Form. Académica do Pai       0.000000
Nacionalidade da Mãe         0.000000
Naturalidade da Mãe          0.000000
Sit. Emprego da Mãe          0.347222
```

Figure 11 – List of relevant variables with missing values

Given these are categorical variables, the methodology used to fill the missing

values was by their mode.

3.1.3. New variable creation

An auxiliary variable – *Age* – was created, matching the different schooling years and the age the student would be on 31st December, not having he or she an irregular academic year. A lost academic year is considered when students fail a school year, when they are transferred to another school or another country, or when they change courses and are obliged to repeat the same schooling year.

For instance, if a student is in the 9th grade and did not lose a school year for failing a year nor for losing one year due to transferences, he or she should be 14.

```
print(dataset['Ano'].value_counts())
def age_age(x):
    if x== '9º ano':
        y= '14'
    else:
        if x== '10º ano':
            y = '15'
        elif x=='11º ano':
            y = '16'
        else:
            y = '17'
    return y
dataset["Age"] = dataset["Ano"].apply(lambda x: age_age(x))
dataset['Age'] = dataset['Age'].astype(int) #convert Age from object to int
```

Figure 12 – Series of unique values of the variable Ano

The new variable itself – *Irregular school years* – is obtained by the subtraction of the new auxiliary variable to the student's age on 31st December.

Having the parents' zip code split into two different variables, a concatenation was made to make so it could be converted into a single column. Then, a binary variable – *Parents Situation* – was created, by comparing the parents' concatenated zip code. The new variable aims to infer rather the students' parents live or do not live together. Despite the fact that multiple missing values were identified, as most parents live together it was considered the same household for the couple.

3.1.4. One-hot encoding

Due to some algorithm's inability to directly work with categorical data, one-hot encoding – a broadly used technique in Machine Learning – was applied to all the textual variables, such as the naturalness and the nationality ones, along with the students' parents' level of education, so that the models could be applied.

One-hot encoding is applied when there is a sequence classification type problem

and categorical data must be converted into numbers through the representation of categorical variables as binary vectors. In technical terms, each integer value is represented as a binary vector composed solely for zero values, except the index of the integer, that is marked with a '1'. For the student below, for instance, the variable *Ano* is represented as the vector (0,0,1,0).

Ano_10	Ano_11	Ano_12	Ano_9
0	0	1	0

Figure 13 – One-hot encoding vector, representing one students' schooling year

3.2. Data analysis

3.2.1. Analysing the variables' correlation

It is fundamental to conduct an analysis regarding the variables' correlation, as it is one of the assumptions for most machine learning algorithms.

In short, correlation designates the statistical relationship between two variables. It can be negative – if they behave in the opposite direction; neutral – if the variables are not related – or positive – if they behave in the same direction.

To assess all the features' dependency, using Python's capabilities – namely the corr() function –, the graphic below was obtained:

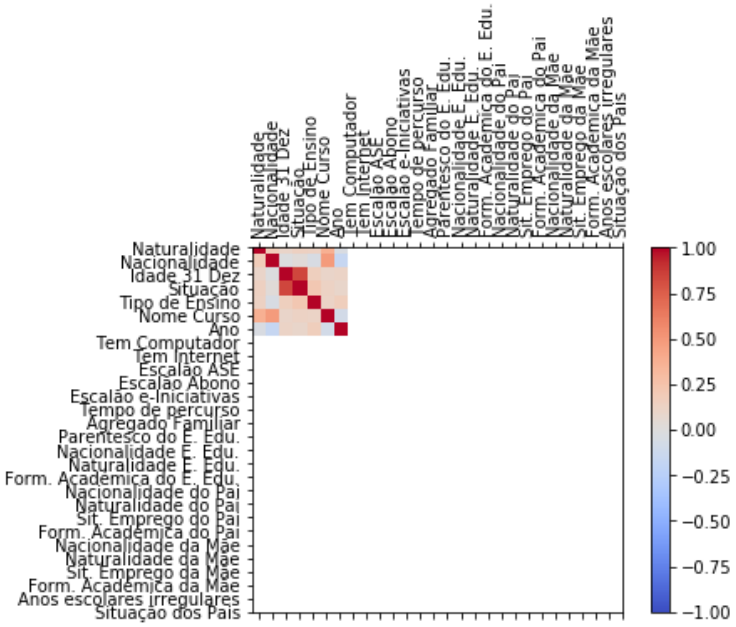


Figure 14 – Grid representing the correlation between the variables

To the royal blue squares would be associated a perfect negative linear correlation; to the red squares would be associated a perfect positive correlation and, to the grey squares it would be associated a neutral dependency between the variables.

For example, the student's age does not provide information on whether he or she has a computer, but it is highly correlated to the students' situation when it comes or not to drop out school – which is the dependent variable. However, as there is no major correlation between this subset of variables, considering no variable might be used to predict another, the multicollinearity assumption is respected. Had there been two variables so correlated that one can be used to predict the other one –multicollinearity –, the models' performance could be affected, and the feature selection would be demanded.

3.2.2. Analyzing variable's importance

A heatmap was drawn to equally illustrate the relevance of the variables considered, being highlighted 5 most significate variables, disregarding *Situação* as it is the dependent one.

Heat maps are an increasingly popular data visualization tool in scientific disciplines, which provide both rich and accessible representations of dynamic processes, easing the visualization and understanding of the dataset at a glance, through a data-driven graphical representation, by using a warm-to-cool color spectrum.

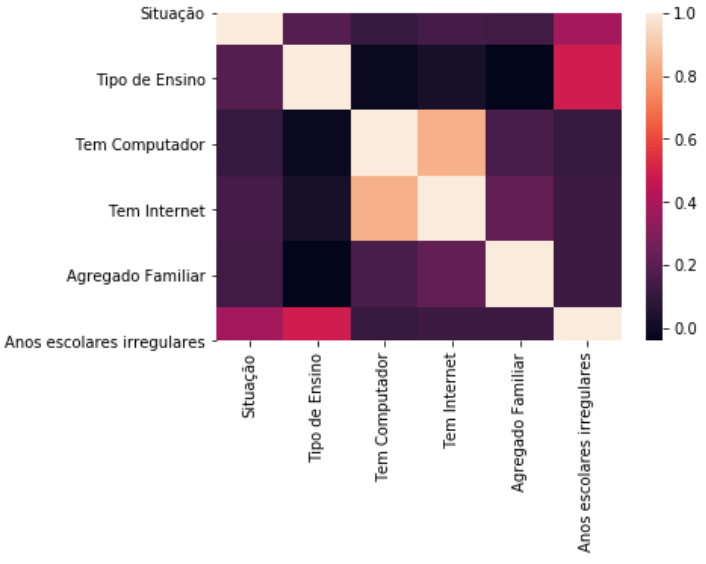


Figure 15 – Heatmap regarding the main variables

The heatmap is a divided grid that, by assigning each value to a color representation, shows the relative intensity of values captured by the eye tracker.

As depicted, the "hotter" the color squares, the higher the intensity of values captured and, on the other hand, the "cooler" the color squares, the lower the value.

Briefly, this tool gives hints on what variables to highlight in the analysis. Particularly, *Tipo de Ensino, Tem computador, Tem Internet, Agregado Familiar, Anos escolares irregulares.*

3.3. Preparing the models

3.3.1. Splitting the dataset

After data pre-processing and preceding Machine Learning algorithms' application, data must be split into two subsets: training – to fit the models in – and testing data – to make predictions on the dataset. The proportion used was 80-20, for training and test set, respectively.

3.3.2. Oversampling and under sampling

Machine Learning Algorithms' main mission is to both enhance and reduce error and their performance are as good as the dataset is balanced.

However, as the histogram below suggests, this is a highly skewed and imbalanced dataset, with a biased class distribution, with solely 10% of students dropping out.

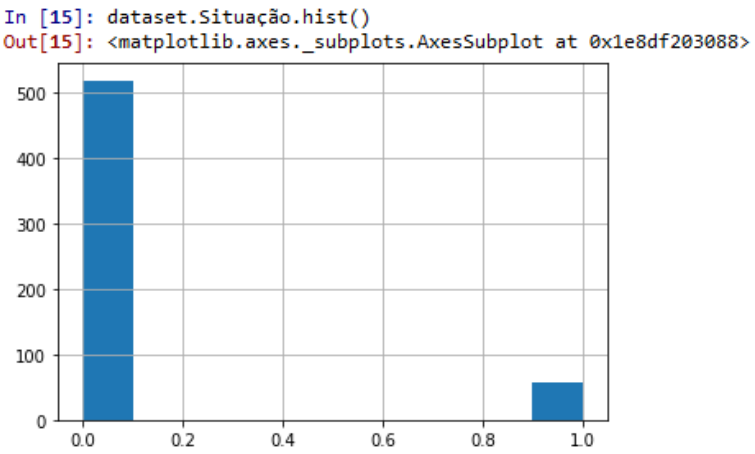


Figure 16 – Distribution of the dataset-dependent variable

It is, then, important to mitigate this skewness and disproportion ratio of observations that can influence machine learning algorithms to further apply, especially regarding the minority, which is the scope of the present analysis.

As the minority is underrepresented and the majority is overrepresented, two naïve random resampling strategies will be carried on, particularly oversampling and under sampling. These can be applied alone or combined; being the latter option, the best one.

This shift in the class distribution will only be applied to the training dataset – to impact the fit of the models –, not to influence the models both outcome and performance. As seen, the distribution in the training set (80% of the dataset) is largely disproportional:

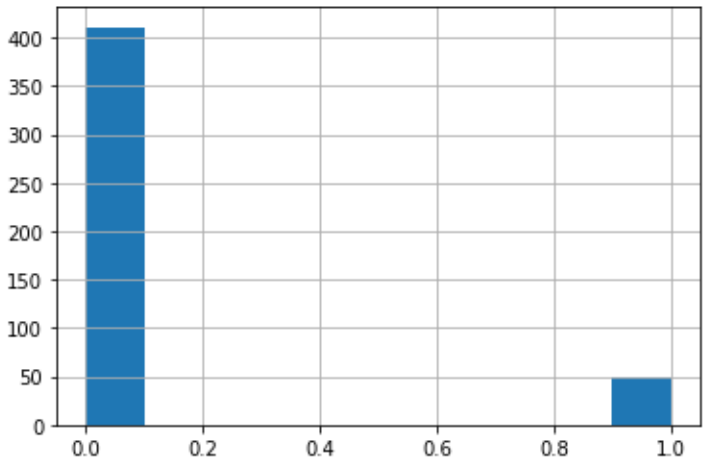


Figure 17 – Distribution of the training set-dependent variable

Starting with oversampling, that will randomly select and duplicate examples in the minority class, i.e. students who do not drop out, the disproportion is going to be evened out:

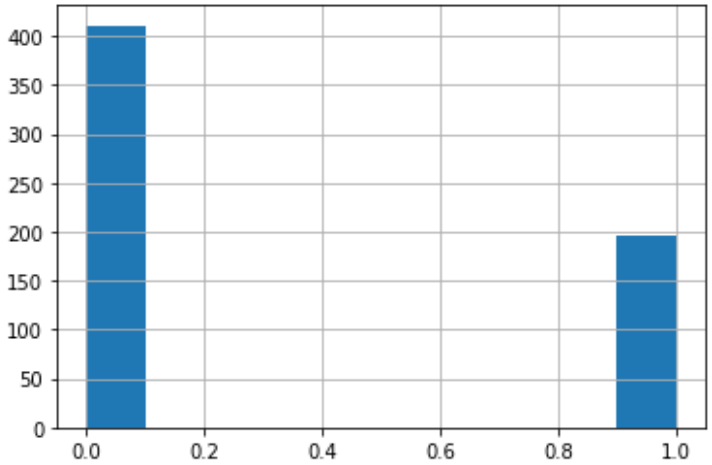


Figure 18 – Distribution of the dataset-dependent variable, after oversampling

Additionally, applying under sampling to randomly select delete examples in the majority class, i.e., students who do not drop out, the sample is going to be more balanced:

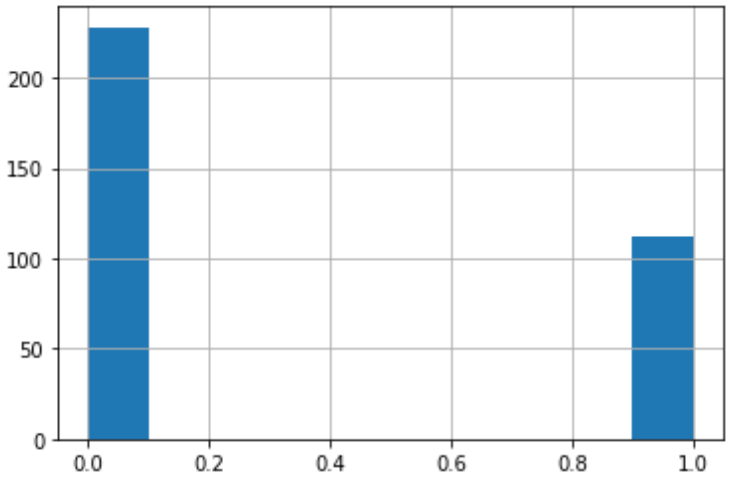


Figure 19 – Distribution of the dataset-dependent variable, after oversampling and under sampling

These operations can be repeat as many times as desired to recreate a new version of the dataset with a more proportional and balanced distribution. It is then reasonable to say the dataset is prepared to be tested.

4. MACHINE LEARNING ALGORITHMS

At a first instance, prior to training the model, DummyClassifier was used to predict how misleading accuracy can be, even without training the model, its' performance is already great, as seen below:

```
Unique predicted labels:  [0]
Test score:  0.9224137931034483
```

Figure 20 – Models' performance without having it trained

This might be an indicative sign that this is not the right tool to check a models' performance.

4.1. AUC-ROC Curve as a performance measurement

The AUC-ROC Curve (Area Under the Curve – Receiver Operating Characteristics Curve) is a largely relevant performance measurement tool for classification problems.

This evaluation metrics is composed by a probability curve – ROC curve – and a degree of separability – AUC –, that measures the models' capability to distinguish between classes.

The confusion matrix below is helpful to understand the following concepts regarding the AUC-ROC curve:

		ACTUAL	
		Positive	Negative
PREDICTED	Positive	**True Positive**	**False Positive**
	Negative	**False Negative**	**True Negative**

$$TPR = Recal = Sensibility = \frac{TP}{\mathrm{TP} + \mathrm{FN}}$$

Sensibility measures the proportion of observations that were correctly predicted to be positive out of all positive observations.

$$Specificity = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

Specificity measures the proportion of observations that were correctly predicted to be negative out of all negative observations.

$$FPR = 1 - \text{Specificity} = \frac{\text{FP}}{\text{TN} + \text{FP}}$$

False positive rate measures the proportion of observations that are incorrectly predicted to be positive out of all negative observations.

Sensitivity and sensibility are inversely proportional, and the ROC curve shows the trade-off between them, by plotting with TPR (True positive rates) on y-axis, against FPR (False positive rates) on the x-axis.

For instance, if the threshold is increased, so is the specificity, as more negative values are obtained. However, if the threshold is decreased, so is the sensibility, as more positive values are obtained.

The closer the AUC is to 1, the better the model's measure of separability and the better its capacity to predict the classes correctly, i.e. '0' as '0' and '1' as '1'.

At a 0.5 threshold, if both curves (positive class: students drop out; negative class: students do not drop out) do not overlap, the model as an ideal measure of separability and perfectly distinguishes positive and negative classes (Sarang, 2018):

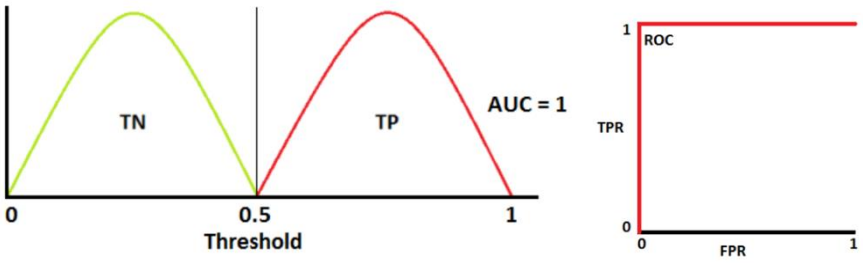


Figure 21 – Model's perfect predictive capacity

When distributions overlap, there are two types of errors (false negative and false positive) that can be maximized or minimized, depending on the threshold. In this case, there is 70% probability that the model will correctly assign positive and negative classes (Sarang, 2018):

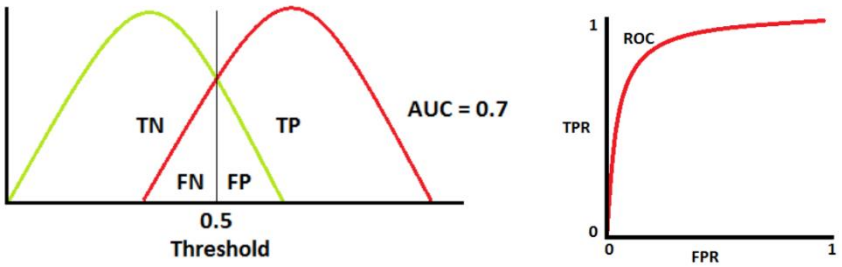


Figure 22 – Model's high predictive capacity

Lastly, when the distributions totally overlap, the AUC = 0.5 and the model is completely incapable of distinguishing the classes. It is predicting positive classes as negative and vice-versa (Sarang, 2018):

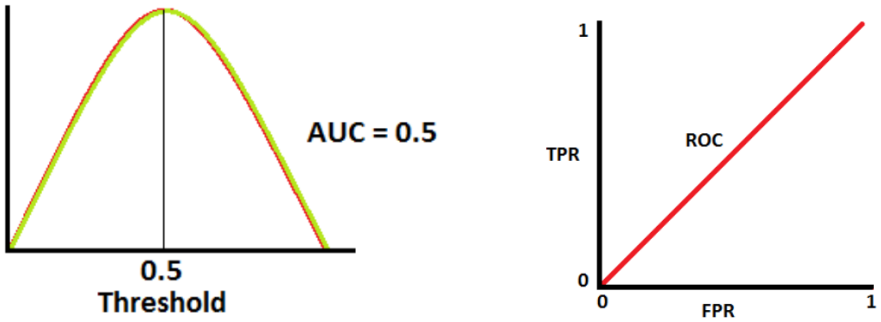


Figure 23 – Model's null predictive capacity

Basically, as closer to the top-left corner, the better the performance; the closer to 45-degrees diagonal the ROC space, the less accurate, as FPR=TPR.

4.2. Logistic Regression

4.2.1. Model's theoretical explanation

Unlike linear regression – used to identify a linear relationship between a target or dependent variable (y) and one or multiple predictors or independent variables (x) –, Logistic Regression aims to make predictions in a scope in which the dependent variable is categorical. Statistically, it tries to estimate a multiple linear regression function, defined as:

$$\text{Logit(p)} = \log(\frac{\text{p(y=1)}}{\text{1}-\text{(p=1)}}) + \beta 0 \ + \ \beta 1 \text{xi1} \ + \ \beta 2 \text{xi2} + \ldots + \beta \text{pxip},$$

where i= 1,...,n.

Within this topic, considering there are two possible outcomes for the target variable – '1' if the student drops out, '0' if the student does not drop out –, Binary Logistic Regressions will be applied, and some assumptions must be considered, namely:

- The dependent variable must have a dichotomous nature, in this case 'dropout' or 'does not dropout';
- There must not be outliers in the dataset;
- There must not be multicollinearity – high correlation coefficients (β) – among the predictors. Duly, if, as checked previously, the variables' relation does not exceed 0.90 the assumption is respected (Tabachnick and Fidell, 2013).

4.2.2. Model Results

The ROC curve, for Logistic Regression is 88%. This means this model has 88% chance to correctly predict positive classes as so and negative classes as so.

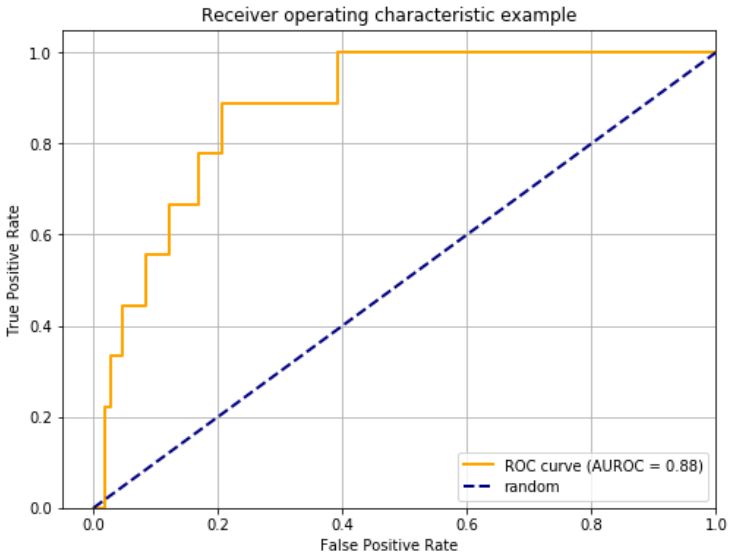


Figure 24 – Logistic Regression AUC-ROC Curve

4.3. Random Forest

4.3.1. Model's theoretical explanation

Random Forest is typically one of the most successful and easy to use supervised Machine Learning algorithms.

It builds and merges an ensemble of decision trees that improve the overall result, by adding diversity to the model. While splitting a node, the best – and not the most important – feature is selected and, as decision trees are added, so is the model's randomness.

Unlike with decision trees – that make predictions by considering previous features and labels, through rules' elaboration –, the random forest algorithm randomly selects observations and features to build decision trees, finally averaging the results, by combining the subtrees.

By solely considering a random subset of features, Random Forest is likely prevent the models to overfit, thanks to the random forest classifier, but are costly in terms of the computation's velocity.

4.3.2. Model Results

The ROC curve, for Random Forest is 77%. This means this model has 77% chance to correctly predict positive classes as so and negative classes as so:

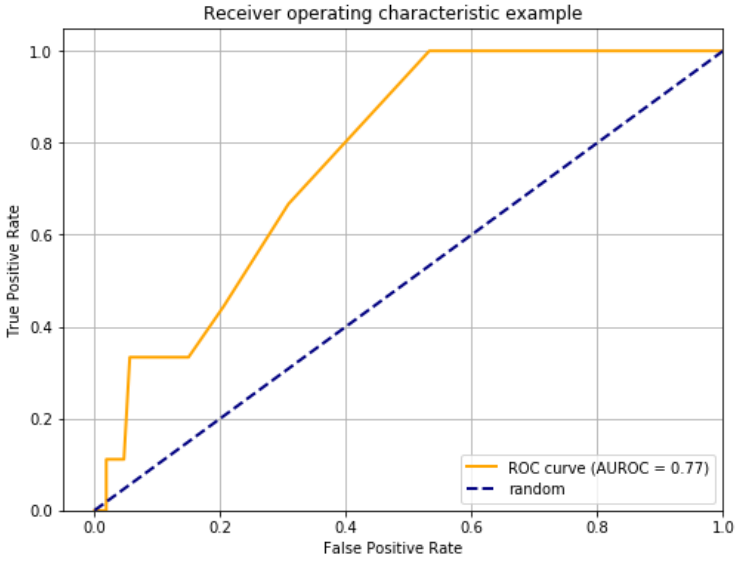


Figure 25 – Random Forest AUC-ROC Curve

4.4. Choosing the best model

The AUC-ROC curve is the performance measure used to evaluate the models' capability to correctly predict the classes.

Knowing that the higher the AUC-ROC area, the better the models' chances to predict positives as positives and negatives as negatives. The model that best does so is the Logistic Regression, that is going to be the elected model to predict whether a student is or is not going to drop out school.

5. CONCLUSIONS, LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORKS

The familiar context – which does include wealth and heath factors, household's level of education, among others – influences one's capacity and conditions inherent to the learning process.

The community where students live dictates the quality of the institution they attend and its resources, their neighbourhood and, hence, their peers; and their likeliness to experience out-of-school educative or cultural experiences.

The students' disengagement with school – that can be noted through the lack of interaction with students' peers or by simply not doing the homework – is a 'red flag'. In fact, it might be reflected by poor school achievements, which trigger the snow-ball-effect and contribute to one's withdrawal.

The early leaving school is damaging for the student him or herself as it conditions him or her socioeconomic life, but also for Government's, in the sense that a decreasing revenue and productivity are regarded.

To understand the expected output is crucial when considering the dataset at first. Then, so that it could be possible and suitable for the Machine Learning algorithms application, it is key to "clear" the dataset, to deal with missing values considering the variables' distribution, to create new variables and to identify trends as well as to understand the variables' correlation and its respective importance. Furthermore, for the school in analysis, that is Escola Básica 2/3 Dr. Azevedo Neves, students of professional courses that do not have computer nor access to the Internet, students whose household is larger and who have had more than one irregular schooling year are more prone to drop out school precociously.

The final step is to train the Machine Learning Algorithms and to choose the one with the highest performance. Out of the two algorithms applied, the Logistic Regression is the one with the highest performance and so, it is the one that should be considered when predicting whether a student is or is not going to drop out, having into account his or her characteristics.

The major limitation is the fact that the variables considered were solely at the student's scope and did not contemplate the student's relationship with the institution

and its representatives, nor with their families and the involving community. Also, had the student's academic performance been included in this analysis, the models could have higher ROC curve rates.

6. BIBLIOGRAPHY

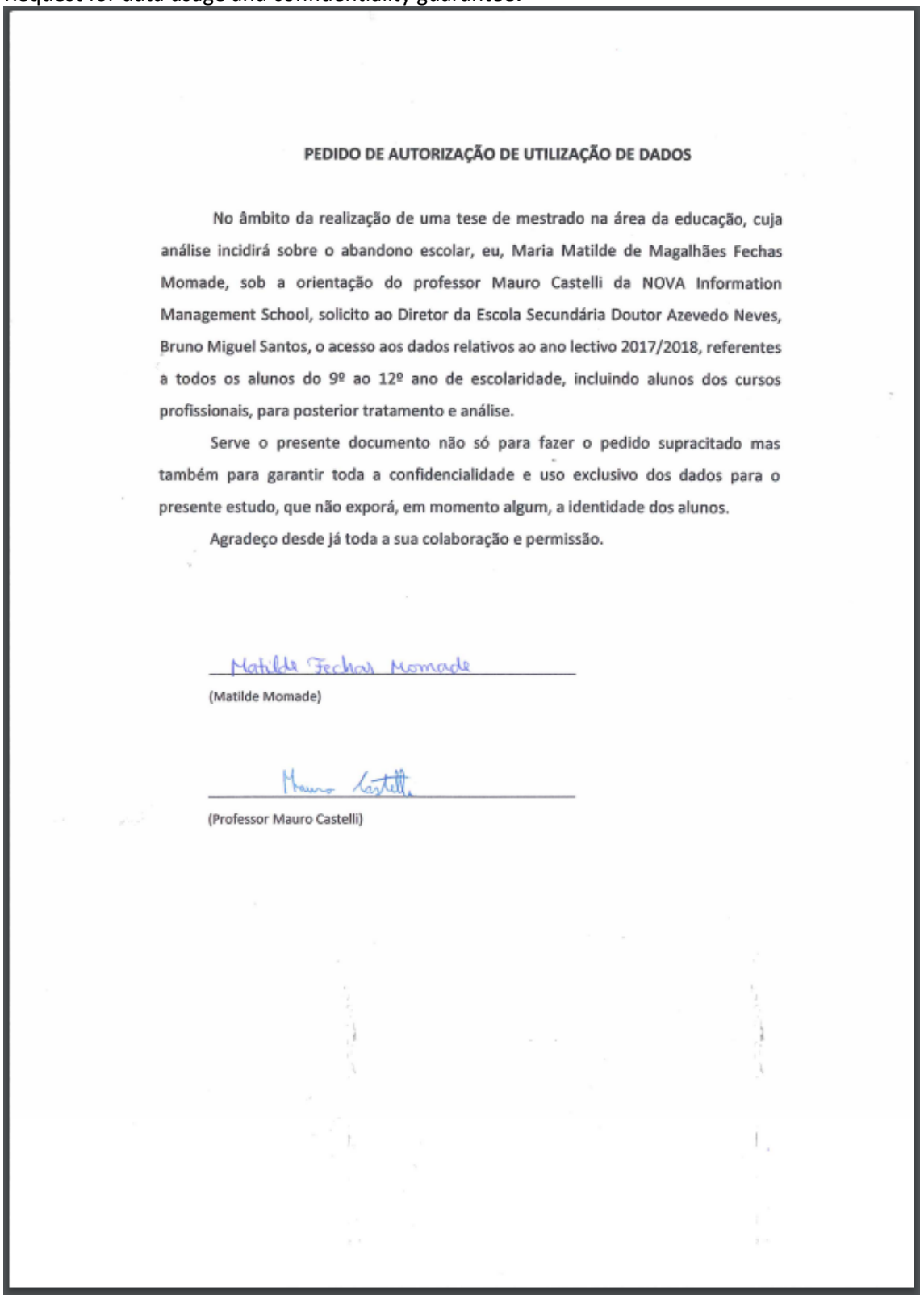
- Lobo, Andreia. (2016). *"Há uma relação muito forte" entre sucesso escolar e condições das famílias* (https://www.educare.pt/noticias/noticia/ver/?id=116082)
- Direcção-Geral de Estatísticas da Educação e Ciência. (2016). *Desigualdades Socioeconómicas e Resultados Escolares – 3º Ciclo do Ensino Público Geral* (http://www.dgeec.mec.pt/np4/97/%7B$clientServletPath%7D/?newsId=147&fileName=DesigualdadesResultadosEscolares.pdf)
- Casanova, Fátima. (2016). *Estudo liga sucesso escolar ao nível socioeconómico e habilitações das mães* (https://rr.sapo.pt/2016/02/24/pais/estudo-liga-sucesso-escolar-ao-nivel-socioeconomico-e-habilitacoes-das-maes/noticia/47731/)
- Marques Costa, Rita and Lusa. (2018). *Há menos abandono escolar precoce, mas 12,6% dos jovens ainda deixam a escola demasiado cedo* (https://www.publico.pt/2018/02/07/sociedade/noticia/ha-menos-abandono-escolar-precoce-mas-126-dos-jovens-ainda-saem-demasiado-cedo-da-escola-1802318?fbclid=IwAR2N3jcx16KOy-S4pEaqUsG2WhelXoqWMm8ilkFiP5BTTjO3WlD2tnVkkXI)
- Jornal de Notícias. (2018). *Amadora tem escola mais "africana da Europa" e das melhores em Portugal* (https://www.jn.pt/nacional/interior/amadora-tem-escola-mais-africana-da-europa-e-das-melhores-em-portugal-9856477.html)
- Belfield, C. & Levin, H. M. Eds. (2007). *The price we pay: Economic and social consequences of inadequate education*. Washington, D.C.: Brookings Institution Press.
- Rumberger, Russell W. (2001). *Why Students Drop Out of School and What Can be Done* (https://escholarship.org/uc/item/58p2c3wp)
- Finn, J.D. (1989). Withdrawing from school. Review of Educational Research, 59, 117-142.
- Jason W. Osborne and Elaine Waters (2002). *Four Assumptions of Multiple Regressions that Researchers should always Test. J. of Practical Assessment, Research, and Evaluation*.

- Naylor, M. (1989). Retaining at-risk students in career and vocational education. ERIC
- Digest No. 87. Columbus, OH: ERIC Clearinghouse on Adult, Career, and Vocational Education. (ERIC Document Reproduction Service No. ED 308400)
- Heckman. (2008). Schools, Skills, and Synapses
- Schoon, Ingrid; Duckworth. (2010). *Leaving School Early – and Making It! Evidence From Two British Birth Cohorts* (https://econtent.hogrefe.com/doi/abs/10.1027/1016-9040/a000063?journalCode=epp)
- Brownlee, Jason. (2017). How to One Hot Encode Sequence Data in Python (https://machinelearningmastery.com/how-to-one-hot-encode-sequence-data-in-python/)
- Vickery, Rebecca. (2019). An Easier Way to Encode Categorical Features (https://towardsdatascience.com/an-easier-way-to-encode-categorical-features-d840ff6b3900)
- Farnsworth, Bryn. (2019). How to Analyze and Interpret Heat Maps (https://imotions.com/blog/analyze-heat-maps/)
- Yiu, Tony. (2019). Understanding Random Forest (https://towardsdatascience.com/understanding-random-forest-58381e0602d2)
- Breiman, Leo; Cutler, Adele. Random Forests (https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm)
- Mao, Wenji; Wang, Fei-Yue. (2012). Cultural Modeling for Behavior Analysis and Prediction (https://www.sciencedirect.com/topics/engineering/random-forest)
- Donges, Niklas. (2019). A Complete Guide To The Random Forest Algorithm (https://builtin.com/data-science/random-forest-algorithm)
- Brownlee, Jason. (2020). Random Oversampling and Undersampling for Imbalanced Classification (https://machinelearningmastery.com/random-oversampling-and-undersampling-for-imbalanced-classification/)
- Brownlee, Jason. (2020). Combine Oversampling and Undersampling for Imbalanced Classification (https://machinelearningmastery.com/combine-oversampling-and-undersampling-for-imbalanced-classification/)

- Boyle, Tara. (2019). Dealing with Imbalanced Data (https://towardsdatascience.com/methods-for-dealing-with-imbalanced-data-5b761be45a18)
- Narkhede, Sarang. (2018). Understanding AUC - ROC Curve (https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5)
- Chan, Carmen. What is a ROC Curve and How to Interpret It (https://www.displayr.com/what-is-a-roc-curve-how-to-interpret-it/)
- Swaminathan, Saishruthi. (2018). Logistic Regression — Detailed Overview (https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc)
- Statistics Solutions. What is Logistic Regression? (https://www.statisticssolutions.com/what-is-logistic-regression/)

7. ANNEXES

Request for data usage and confidentiality guarantee:



PEDIDO DE AUTORIZAÇÃO DE UTILIZAÇÃO DE DADOS

No âmbito da realização de uma tese de mestrado na área da educação, cuja análise incidirá sobre o abandono escolar, eu, Maria Matilde de Magalhães Fechas Momade, sob a orientação do professor Mauro Castelli da NOVA Information Management School, solicito ao Diretor da Escola Secundária Doutor Azevedo Neves, Bruno Miguel Santos, o acesso aos dados relativos ao ano lectivo 2017/2018, referentes a todos os alunos do 9º ao 12º ano de escolaridade, incluindo alunos dos cursos profissionais, para posterior tratamento e análise.

Serve o presente documento não só para fazer o pedido supracitado mas também para garantir toda a confidencialidade e uso exclusivo dos dados para o presente estudo, que não exporá, em momento algum, a identidade dos alunos.

Agradeço desde já toda a sua colaboração e permissão.

Matilde Fechas Momade

(Matilde Momade)

Mauro Castelli

(Professor Mauro Castelli)