INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Student Data Prediction

Nuno Miguel Soares Fialho de Carvalho

Master in Computer Engineering

Supervisor:
PhD Elsa Alexandra Cabral da Rocha Cardoso, Assistant Professor,
Iscte – Instituto Universitário de Lisboa

Co-Supervisor:
PhD Fernando Manuel Marques Batista, Associate Professor,
Iscte – Instituto Universitário de Lisboa

October 2021

TECNOLOGIAS
E ARQUITETURA

Department of Information Science and Technology

Student Data Prediction

Nuno Miguel Soares Fialho de Carvalho

Master in Computer Engineering

Supervisor:
PhD Elsa Alexandra Cabral da Rocha Cardoso, Assistant Professor,
Iscte – Instituto Universitário de Lisboa

Supervisor:
PhD Fernando Manuel Marques Batista, Associate Professor,
Iscte – Instituto Universitário de Lisboa

October 2021

To my wife Bela and daughters Sofia and Sara, for their unquestioning love, unwavering support, encouragement, patience, and assistance in overcoming the challenges that arose along the way.

Acknowledgment

This document on data-driven analysis of Dropout comes from the author's work at ISCTE - Instituto Universitário de Lisboa's soft skills programs. However, it would not have been possible without the help of a number of people and organizations to whom we express our heartfelt gratitude:

For the opportunity provided by ISCTE - Instituto Universitário de Lisboa;

I am grateful to Professors Elsa Cardoso and Fernando Batista, who served as co-supervisors on this project, for their encouragement, knowledge sharing, and significant contributions to this dissertation. Above all, thank you for accompanying me on this journey.

To the professors at ISCTE - Instituto Universitário de Lisboa, for their valuable contributions to my scientific education and my academic path;

To my family and colleagues for their help, support and knowledge.

Thank you all from the bottom of my heart!

Resumo

Um dos grandes desafios para a educação é de ser capaz de oferecer programas de ensino à distância onde os estudantes possam sentir que são uma mais-valia para a sua formação académica. Embora o desenvolvimento tecnológico torne possível ultrapassar as barreiras físicas de uma sala de aula e desta forma alcançar muito mais estudantes, há, ao mesmo tempo, uma maior dificuldade em fazer com que a oferta tenha a qualidade espectável. A falta de seleção de candidatos, a adequação dos programas ao ensino à distância ou a falta de interação entre estudantes e entre estudantes e docentes são fatores que contribuem para taxas de abandono escolar nestes cursos de ensino à distância.

O objetivo deste documento é tentar compreender quais os fatores que mais contribuem para as taxas de abandono escolar, como identificar antecipadamente os alunos em risco de abandono escolar, e como agir de forma a diminuir este risco.

Para tal, utilizaremos dados de programas à distância de *soft skills*. É feita uma análise exploratória dos dados para compreender quais os fatores que mais contribuem para esta taxa de abandono escolar, são aplicados algoritmos de aprendizagem automática para classificar os estudantes em risco de abandono escolar, sendo assim possível identificar estes estudantes com antecedência e promover ações para evitar estas desistências.

Palavras-chave: Ensino à distância, Aprendizagem Automática, Previsão de Desistências, Fatores de Desistência.

Abstract

One of the great challenges for education is to be able to offer distance learning programs where students feel that these programs are an added value to their academic training. Although technological development makes it possible to overcome the physical barriers of a classroom and thus reach many more students with this distance learning offer, there are at the same time a greater number of difficulties in reaching them with the proper quality. The lack of selection of candidates, the suitability of the programs for distance learning or the lack of interaction between students and between students and teachers are factors that contribute to high dropout rates of students in these distance learning courses.

The purpose of this dissertation is to figure out which factors contribute the most to dropout rates, how we can identify in advance students who are at risk of dropping out, and how we can act to decrease this risk.

To do this, we will use data from distance programs of soft skills where an exploratory analysis of the data will be done to understand which factors contribute most to this dropout rate and where machine learning algorithms will be applied to classify the students at risk of dropping out, thus being possible to identify these students in advance and promote actions to avoid these dropouts.

Keywords: Distance Learning, Automatic Learning, Dropout Forecasting, Dropout Factors

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Although distance learning is not a new phenomenon of this millennium, the emergence of the so-called web 2.0 has opened up new possibilities in a variety of fields, including education.

The popularity of online education has risen in recent years, owing largely to MOOCs (Massive Open Online Courses), but while the number of students choosing this path is growing, dropout rates are also rising. Students are less concerned about the courses they enroll in and for which they will have, or not, aptitude as access becomes easier and access conditions become less controlled. Student retention has always been an issue in distance learning, and the concern is growing as the number of students enrolling in and dropping out of these courses increases.

It is necessary to adapt the courses to the characteristics of distance learning and teaching methods [30].

Some changes must be made also from the students' perspective. Students have difficulty fully engaging in the online course and learning from it [37].

The distinction is such that calculating what is a success or what is considered a failure requires a different perspective. In contrast to traditional education, not all students who do not complete the whole program or who do not request a certificate of completion of the course are considered dropouts. In [16] can be found that in traditional education, the success rate is the number of certificates divided by the number of enrollments and the remaining is considered as dropout. In contrast on online education, this rate is not calculated in this way, since some students enroll on online courses with the intention of attending part of the program. The rate must consider the students' expectations and involvement. A student who enrolls in a course with the intention of taking only a portion of it and completes that portion, should not be considered a dropout because the student achieves his objectives.

The goal of this research is to determine what constitutes a dropout in online education, identify the main factors that contribute to this dropout, and develop a predict model that can inform if a student is at risk of dropping out at an early stage, allowing

time to take preventative measures.

1.1 Motivation

Education institutes can reach a much larger number of students by offering programs in online education. This has the potential to alter the way institutions work and open up plenty of new possibilities. Online education, on the other hand, varies greatly from traditional education in a variety of ways. If the programs reach a larger audience, the students who enroll have even more disparities in terms of educational backgrounds and levels.

The concerns surrounding the student dropout rate are something that both conventional and online research have in common. Even though it is still a concern in both conventional and online learning, it occurs at a much greater level in online learning, exacerbating the problem.

Typically, a students' decision to drop out is not made in a single moment. Normally, it is a problem that worsens over time, culminating in one drop over the top and a decision to dropout.

We can try to predict if a student has a dropout intention at an early stage of the process by analyzing student characteristics and behavior throughout the program. By obtaining an early forecast, some steps can be taken to persuade the student to abandon his or her plans to drop out and continue with the program.

1.2 Background

COVID-19 has had a significant impact on the world since the late 2019 and early 2020. Lockdown is a tool used by many countries in order to contain the virus. These constraints have increased the demand for online education. This type of education had been growing in popularity even before COVID-19, but the virus accelerated it even further. Online education reaches many more students, but due to its widespread use and differences from traditional education, it faces numerous challenges.

Dropout rates are one of the most serious problems in online education.

Student dropout results in a large number of students enrolling in the program but a small number of students completing it. As stated in [37], this increases the difficulty of teaching with quality, as well as designing and implementing programs.

Dropout rates is an issue in traditional education, but it is even bigger problem in online education. Nonetheless, the use of Learning Management Systems (LMS) in online education can provide us with measurable features on student interaction and engagement, as well as data to better predict students' loss of engagement, which leads to a dropout intention.

1.3 Research Questions

We intend to analyze how students' and educational institutions' commitment can be improved in this dissertation, and as such, we will attempt to answer the following questions:

- What are the most effective characteristics for predicting student dropout in online education?
- Which algorithm provides us with better results in predicting student dropout in online education?
- How can we develop an early warning system for teachers regarding student dropout in online education?

1.4 Objectives

Soft skills programs are lectured in blended learning and the online section is delivered through an online distance learning platform at Iscte - Instituto Universitário de Lisboa. The goal of this project is to identify students in these soft skills programs who may be on the verge of dropping out, and to develop an alert system that will allow intervening with these students and attempting to change their minds.

For a better design of the program and modules, it will be necessary to have a number of students that is as stable as possible. Anything that can be done to avoid dropouts will be an improvement for the program. On the other hand, since the dropout rate is one of the biggest complaints about distance learning, by lowering the number of dropouts, we will improve one of the biggest problems of distance learning.

In the first phase, we will collect data from information systems to determine which characteristics have the greatest impact on student dropouts. Following the identification of these characteristics, several machine learning algorithms will be evaluated in order to determine whether a student intends to drop out as accurately as possible.

After obtaining the best result among the various algorithms, we should fine-tune the one that offers us the best results, in order to obtain the best possible result from the data.

With the algorithm running, the students marked as tending to drop out should be identified, so that actions can be taken to prevent this tendency.

1.5 Research Method

The CRIP-DM method will be used to further the research. It is appropriate for the process to be implemented, as the name suggests (CRoss Industry Standard Process for

Data Mining - CRISP-DM).

Figure 1: CRISP-DM Process Diagram.

The following procedures must be followed in order to use the CRISP DM:

- Business Understanding - The Business Understanding phase focuses on comprehending the projects' goals and requirements. The objectives and scope must be defined during this phase.
- Data Understanding - During this phase, the focus should be on the data obtained, with an attempt to investigate the data and its quality.
- Data Preparation - The data must be prepared before it can be used. For the model to work, the attributes must be chosen, cleaned, and formatted.
- Modeling - During this phase, we must select the algorithms to be used, construct the model to be implemented, and interpret the results.
- Evaluation - After the data models have been run, they must be evaluated and adjusted to produce results that meet the defined success criteria. All of these steps should be revised if the data does not agree.
- Deployment - If the process is in compliance, the model should be put into production so that the results can be seen.
- Even so, the model should be evaluated continuously and, if necessary, adjusted on a regular basis and rephrasing the previous steps.

1.6 Document Structure

The remainder of this document has been organized as follows: Chapter 2 is about the work done in this area; On Chapter 3 is the data gathering process and understanding; Chapter 4 is about the approaches used to solve the problem and an evaluation and a discussion of the obtained results with the different models; Finally, Chapter 5 presents the major conclusions of this research.

Chapter 2

Literature Review

In this chapter, we will go over the process of researching previous works on topics related to online learning, as well as the reasons for success and failure.

Both students and educational institutions have shown a high level of acceptance for online education. However, after the initial enthusiasm, the programs have had lower and lower success rates, and as a result, a higher and higher dropout rate [16]. Part of the problem is a lack of adequate distance learning means and programs, the absence of minimum admission requirements, and students' lack of preparation, all of which contribute significantly to failure and early withdrawal from the program.

It is also fascinating to see what constitutes success and failure in distance education. According to [16], not all students who do not finish their studies are considered unsuccessful. Some students enroll in a course with the intention of only completing a portion of the programs' modules that interest them. To consider a student as a failure or dropout who never intended to complete the program is not the best approach. In these cases, it would be ideal to conduct a survey of students at the time of enrollment, asking them about their intentions and expectations for the program.

The research objectives are described in Section 2.1, followed by a summary of the research process in Section 2.2. Based on the research process, an analysis of the scientific output is completed in Section 2.3, finishing with a comparison of the findings to the ongoing study in Section 2.4.

2.1 Search Process

The aim of this analysis was to establish a Systematic Literature Review (SLR)[6] on what has been published over the last 5 years on the subject of the critical success factors and the drop-out of online higher education courses. Research was conducted using the Web of Science (WoS) Core Collection. In order to focus on the researh, the topic of the query was "Online Learning Courses Dropout." On the first filter, we chose to only get

results from the last 5 years and in Portuguese or English, as these are languages that do not need external tools to be understood.

213 results were obtained in the first research. The queries have been refined in order to obtain the results with Open Access, to have access to full documents, filtering to 77 results. From this point 2 results with the type of document as a review were excluded, since these are usually studies in previous articles and do not provide new information. Continuing to refine the search, only texts from the "Education Educational Research" or "Computer Science Information Systems" WoS categories were also selected. At this time, we have a total of 51 research found.

The full search query used in the WoS Core Collection is " TOPIC: (online learning courses dropout) AND LANGUAGE: (English OR Portuguese) Refined by: Open Access: (OPEN ACCESS) AND [excluding] DOCUMENT TYPES: (REVIEW) AND WEB OF SCIENCE CATEGORIES: (EDUCATION EDUCATIONAL RESEARCH OR COMPUTER SCIENCE INFORMATION SYSTEMS) Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI, CCR-EXPANDED, IC Timespan=Last 5 years"

As a summary of what has been included and excluded, the criteria used is shown in Table 1.

Table 1: Inclusion and Exclusion Criteria.

Inclusion Criteria	Exclusion Criteria
Open Access	Before 2017
Education Educational Research	Reviews
Computer Science Information Systems	
Language English or Portuguese	
Article and Proceedings Paper	

2.2 Analysis of scientific production

We obtained 51 documents after selecting the documents from the WoS database. In Figure 2, we can analyze their distribution over the years.

As we can see in the graph, the largest number of selected publications has been focused over the last two years and are mostly Journal Papers.

Analyzing in more detail, we can see in the Table 2 the number of citations aggregated by year of publication.

Although the largest number of publications are from 2019 and 2020, there is a higher number of citations for the 2017 publications. More recent publications have, as expected, fewer citations.

Out of the results found, we went to the screening to read the abstract of the documents and to establish a relationship between the articles and what we intend to investigate.

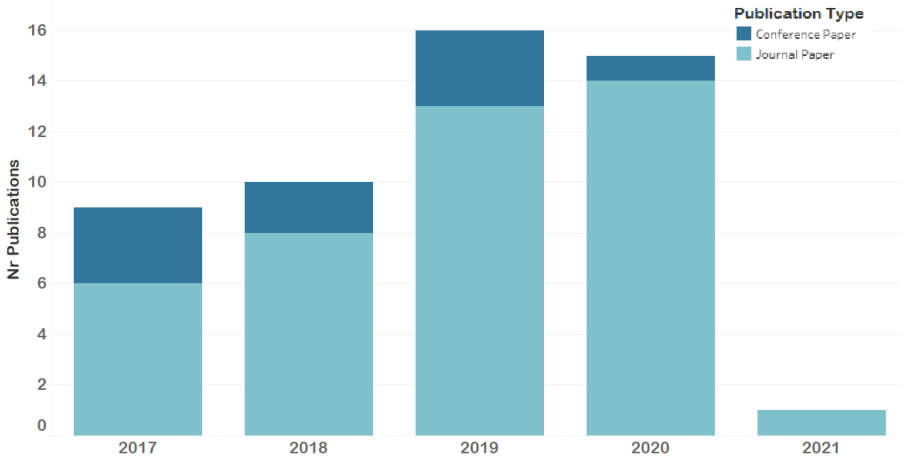

Figure 2: Publications by year.

Table 2: Publications and Citations by year.

Year	Publications	Citations	% Citations
2017	9	143	56%
2018	10	42	16%
2019	16	54	21%
2020	15	16	6%
2021	1	0	0%
Total	51	255	100%

The Table 3 shows us the list of the articles chosen, after eliminating the ones with content far from what we pretend to study.

Table 3: Publications.

Authors and Year	Study and Model	Results
Henderikx et al. [16] 2017	This article aims to define online educations' success and failure, as well as how it differs from traditional education.	Because not all students who enroll in an online program intend to complete it, not all students who do not complete it are considered dropouts. Only those who fail to meet the goals they set for themselves.
		Continued on next page.

Table 3 – continued from previous page.

Authors and Year	Study and Model	Results
Alario-Hoyos et al. [1] 2017	A study of student motivations and learning methods was conducted. It also aims to clarify the various approaches to calculating the dropout rate in an online course.	It is necessary to research students' motivations in order to tailor learning strategies to their needs and thus improve program outcomes.
Sunar et al. [30] 2016	Study of how social relationships in online courses reduce dropout rates and how dropout rates can be reduced by increasing student motivation for the program.	Students who are predisposed to participate in forums and who receive more support from their peers are more likely to complete the program.
Klemke et al. [19] 2018	Increased student engagement is the subject of a study. The flipped classroom concept, gamification, and learning analytics are all used to achieve this goal.	In the context of a flipped MOOC, the balanced application of some gamification concepts brings benefits by using learning analytics to link the information of each student with the overall learning.
Romero-Rodriguez et al. [25] 2019	An investigation into how gamification can boost student engagement in online programs.	Gamification encourages students to compete and interact, which increases their interest in the programs they attend and the outcomes they achieve.
Henderikx et al. [15] 2019	Barriers and success factors in an online course are discussed in this article.	Not all barriers are obstacles to student success; in fact, overcoming some of them can increase motivation and satisfaction in completing the program.

Continued on next page.

Table 3 – continued from previous page.

Authors and Year	Study and Model	Results
Jacobsen [17] 2019	A study of a group of students who dropped out of a previous version of a MOOCs' motivation and study strategies	Some students never had the intention to complete the course, but to access specific contents, and the remaining evaluated students informed that they did not complete the course due to lack of time.
Sukhbaatar et al. [29] 2019	The study employs an artificial neural network with data from the LMS and student grades to identify students at risk of dropping out as early as possible.	The outcomes of after the first quiz, 25 percent of accurate predictions were made, and after the mid-term exams, 65 percent were made. where students on the verge of dropping out were successfully identified.
Henderikx et al. [14] 2018	A study about obstacles that students face when taking online courses.	Barriers to technical and online learning skills, social context barriers, design and management of course expectations, and time, support, and motivational barriers have all been identified.
Haiyang et al. [12] 2018	The study builds a predictive model to identify students at risk of dropping out using a time-series classification method based on student behavioral data and actions.	Later stages have near-90 percent accuracy, but earlier stages have results that are also above 50 percent accurate.
Stathakarou et al. [28] 2018	The impact of branching points on learners' engagement is investigated in this study.	Branching had a negative impact on the completion of the module activities, it was concluded.

Continued on next page.

Table 3 – continued from previous page.

Authors and Year	Study and Model	Results
Xie [35] 2019	Statistical analysis categorizing students according to the data obtained on video viewing.	As students progress through the program, their viewing time and dropout rate decrease (with the exception of the lindy effect).
Wen et al. [34] 2019	Based on learner behavior data from MOOC platforms, this study uses a Convolutional Neural Network (CNN) model to predict dropout.	The final result shows that both precision and accuracy are consistently above 85 percent, indicating an excellent performance in predicting student dropout.
Pilli et al. [23] 2018	This research uses a SWOT analysis to better understand and improve the value that MOOCs can bring to higher education institutions.	As a result, the strengths, weaknesses, opportunities, and threats to MOOCs are identified, allowing them to improve, as well the institutions.
Atapattu and Falkner [2] 2018	A cross-examination of students' interactions with videos and a contextual analysis of the videos' explanation text	The findings show that features of lecturers' video discourse and video interactions have consistent correlations, demanding focus on the peaks of video interaction that could mean difficulties by the students.
Chen et al. [8] 2020	The relationship between students' preexisting misconceptions and retention in a MOOC setting is investigated in this study.	The findings demonstrate that misconceptions among students are a barrier to persistence in MOOCs.
Bozkurt and Akbulut [5] 2019	Study between cultural context and student dropout patterns, using data from the LMS on social data analytics followed by a correlational approach.	Dropout patterns are found to be influenced by cultural context.

Continued on next page.

Table 3 – continued from previous page.

Authors and Year	Study and Model	Results
Thomas et al. [33] 2017	The study was carried out in order to improve student interaction and readiness to take the course, thereby reducing dropout rates.	Formative assessment was shown to be directly connected to attrition in this study; however, including feedback within the assessment reduced mind wandering, improved sense of comprehension, and improved predicted performance.
Guajardo Leal, GonzĆ, et al. [11] 2019	An investigation into the relationship between contextual factors (demographic characteristics), student behavior in the classroom, and learning outcomes. The goal of this research is to figure out what factors contribute to student success.	What factors (contextual and behavioral) most influence school success are identified.
Tahiri et al. [32] 2017	The paper proposes that online courses be redesigned to create a learning path that benefits both individual and collaborative learning.	Individual and collaborative learning progress is improved by forming more homogeneous groups of students.
He et al. [13] 2020	The goal of the study is to use a recurrent neural network on students' biographical and behavioral data to predict their performance during the course.	Early detection of students at risk of dropping out of class allows for early intervention.
Ros et al. [26] 2020	The study examines students' personal perceptions of the use of gamification in a cybersecurity course.	The findings point to a strong link between gamification use and student engagement.

Continued on next page.

Table 3 – continued from previous page.

Authors and Year	Study and Model	Results
Sureephong et al. [31] 2020	The paper investigates the impact of work intensity in an online course on course completion and dropout rates.	The findings suggest that as the workload increases, so does the likelihood of dropping out or transferring to a face-to-face course.
JUNIOR et al. [18] 2019	The goal of this research is to improve the transition from SPOC to MOOC courses in order to increase access and improve success rates.	There is a transition with improved success and dropout rates by conducting surveys to learn about the needs of the students.
Bloemer et al. [3] 2018	Study attempting to develop a model for predicting students' rankings based on their prior grades and biographical information.	Courses that can be a stumbling block for students without a degree have been identified.
Rothkrantz [27] 2017	The paper is an investigation into how to redesign MOOC courses to increase student interest and attention, and thus increase the success rate.	Emotions play a significant role in a students' ability to learn. If it can be positively influenced, the students' interest will rise, and the success rate will rise as well.
Chen et al. [9] 2021	In an online course with frequent formative tasks, quizzes, and tests, researchers looked at student engagement, learning outcomes, and perceptions.	Students' attention and commitment increase as a result of the constant interactions (tasks, quizzes, and formative tests), and their performance improves.
Klemke et al. [20] 2020	An article on how gamification can boost student engagement (especially in MOOCs) and a platform to help with gamification course design	Although there have been some implementation issues on MOOC platforms, gmaification has contributed to more appealing courses for the most part.

Continued on next page.

Table 3 – continued from previous page.

Authors and Year	Study and Model	Results
Lee et al. [21] 2020	Instructors have not given much thought to self-regulated and effective self-study, but it is one of the concerns of course designers.	According to the findings of a multilevel regression, the effectiveness of MOOCs increases with the success of self-control study strategies.
Cagiltay et al. [7] 2020	A study of student characteristics and MOOC success predictors	As predictors of final grades, were identified the variables View course once, view course half, total forum posts, total number of chapters, average number of completed chapters, course length in weeks, and average age of students.
Mourdi et al. [22] 2020	Using machine learning algorithms, the paper attempts to divide students into three groups: successful, unsuccessful, and dropouts. It was used the following six machine learning algorithms: Decision Trees, Random Forest, Ensemble Method, KNN, SVM, Nave Bayes, KNN, SVM, Nave Bayes, KNN, SVM, Nave Bayes	The results obtained have an average accuracy of 92 percent in identifying students. Early action can be taken because identification is done throughout the course.
Zheng et al. [37] 2020	The goal of the paper is to identify students who are on the verge of dropping out. It employs a Convultional Neural Network, as well as feature extraction and time series analysis.	For dropout prediction, the results show an accuracy of 87 percent, as well as precision, recall, and a f1-score of 86 percent.

Continued on next page.

Table 3 – continued from previous page.

Authors and Year	Study and Model	Results
Yin et al. [36] 2020	The goal of the paper is to use machine learning algorithms to predict student dropout. It does so by employing a Convultional Neural Network and data from the 2015 KDD Cup.	They are able to achieve an accuracy of more than 85 percent starting in the second week of the course.
Rabin et al. [24] 2019	Using NLP - n-gram and keyness applied to learning sequences, researchers looked at the learning behavior of students who completed their initial expectations by enrolling in a MOOC versus those who did not.	Students who met their expectations had fewer deviations in their learning path, according to the findings.
Borrella et al. [4] 2019	Aims to identify students at risk of dropping out of MOOCs using Random Forest and Logistic Regression techniques applied to enrollments, grades, time between clicks and clicks on modules, and module material.	Eighty percent of dropouts can be identified using the proposed model.
Gering et al. [10] 2018	A study that uses logistic regression to assess the correlations between 28 variables and student success in online courses in order to determine what causes student success.	The variables that explain success differ depending on the academic level of the students.

2.3 Related work

Several studies have been conducted in response to the huge growth in online learning in recent years, primarily through Massive Open Online Courses (MOOCs). One of the most significant issues raised in relation to MOOCs is their high dropout rate and low success rate. As discussed in articles [5, 14, 15, 16] there are approximately 10% success rates and 90% dropout rates.

To investigate the issue, there has been an approach that highlights the differences between traditional classroom courses and online courses, particularly MOOCs, where the dropout rate cannot be viewed as all students who do not complete the course or do not request a certificate of completion. In a MOOC, there is a significant number of students who enroll with the intention of completing only a portion of the modules or students who lack the academic or social skills to complete the course. This is addressed in articles such as in [1, 15, 16].

Obtaining a prediction of students at likely to dropout is one strategy for minimizing the effects. If students at risk of dropout are identified ahead of time, we can take steps to assist them and, as a result, change their expected dropout predicted path. Machine learning algorithms such as Logistic Regression, Support Vector Machine (SVM), and Decision Trees have been used, but Deep Learning algorithms such as Long Short Term Memory (LSTM) Neural Networks, Recurrent Neural Networks (RNN), and Convolutional Neural Networks have moved a step forward (CNN), achieving better accurate results. We have examples of these approaches in [7, 11, 34, 37]

Despite the fact that the articles use similar algorithms, the way they are approached is different. As an example, there is a article that use Feature-Weighted and Time-Series (FWTS) for a better selection of weighted attributes in CNNs, such as [37], to compensate for the lack of correlation between neighbors attributes, which is common in the CNN algorithm. On other example, in [11] is used a Logistic Regression to the student engagement to predict course completion.

Some use the information based on the behavior from the student with the learning platform like [12] using a time series to predict or like in [29] using Artificial Neural Network, other focus their study on the learner features, the course characteristics and compared to the certification rates like in [7]. Another paper [35], uses a survival analysis focused on the length of a courses' views.

Because the articles found are mostly about MOOCs, and because ISCTEs' online programs are not of the same nature, beginning on not having the same duration, the methods to be used will have to be adapted, because the characteristics of the modules we will obtain in the data set will be different, just as there will be differences in the data we will obtain from the students.

Chapter 3

Data Understanding

The data used in the scope of this thesis concerns data extracted from an online learning system, and complemented by socio-demographic data from students enrolled in these programs. Such data was gently provided by the Information Systems Development Office of Iscte - Instituto Universitário de Lisboa.

Section 3.1 presents an overview of the data and describes the three datasets. Section 3.2 presents and describes the available information (fields) contained in each one of the datasets. Section 3.3 presents a number of calculated fields that were added in order to simplify the analysis. Section 3.4 presents a number of relevant descriptive statistics about the data.

3.1 Data Overview

The Information Systems Development Office of Iscte - Instituto Universitário de Lisboa was tasked to extract data from distance learning program support systems as well as socio-demographic data from students enrolled in these programs.

There were three different data sets provided. They have all been anonymized to prevent any personal data exposure. One containing data from the distance learning systems' log. Another contains information on module requirements as well as student interaction in the modules. Finally, one with the students' socio-demographic information.

The fields *ID Fenix* (Fenix ID) , *ID Curso* (Program ID) , and *ID da Versão do Curso* (Program Version ID) link the logs and programs datasets, while the ID Fenix field can link all three datasets. A data dictionary indicating the correspondence between course, version, and module IDs and the names they stand for was also delivered by the Iscte - Instituto Universitário de Lisboa Information Systems Development Office.

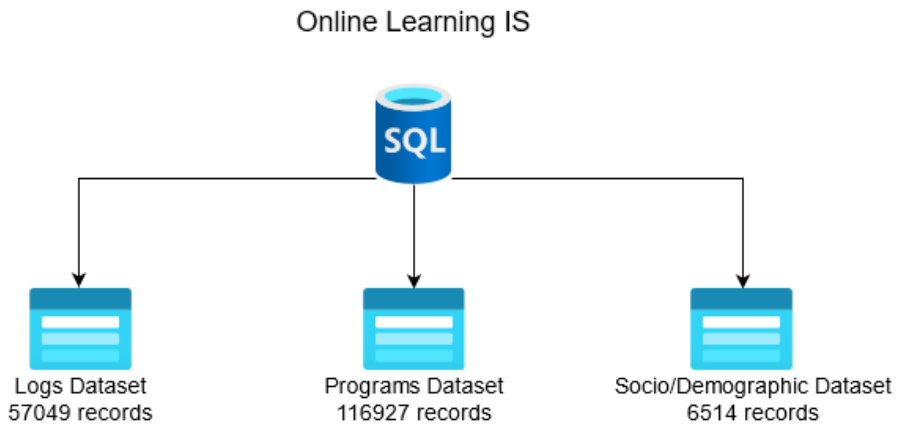

Figure 3: Datasets.

3.2 Data Description

The logs dataset contains information on the students' major actions, since March 2018. The following is a list of the actions kept in the logs:

- *Inscreveu-se no curso* - Program enrollment ;
- *Começou o curso* - Program start;
- *Completou todos os módulos* - All modules completed;
- *Começou o questionário* - Quiz started;
- *Respondeu ao quetionário* - Quiz answered;
- *Completou o curso* - Program completed;
- *Acedeu ao certificado* - Certificate access.

Aside from the list of actions, the dataset has information about the Fenix ID, Course ID, and Course Version ID columns as integer-valued columns that identify the student, course, and version, respectively. The action column is a categorical variable that indicates the action taken by the student in a given course and version. Finally, the actions' date variable indicates when it was created.

A sample of the data from this data set is shown in Table 4.

Figure 4 shows that the months of February, March, and November have traditionally had the most activity, while the months of July, August, and September have had the fewest. We can also confirm that the number of actions has increased in recent years.

The figures in 2021 are still low, but we only have data until the middle of May. With the figures from March and the total of what happened until the 17th of May, it is expected that by the end of the year, they will be higher than in 2020.

Table 4: Logs Dataset sample.

Fenix ID	Course ID	Course Version ID	Action	Date
2963527436533	4	24795	Começou o questionário	2019-06-17
2963527436720	4	24795	Acedeu ao certificado	2021-01-18
2963527436720	4	24795	Respondeu ao questionário	2021-01-18
2963527436720	4	24795	Começou o questionário	2021-01-18
2963528465047	4	24795	Acedeu ao certificado	2021-04-05

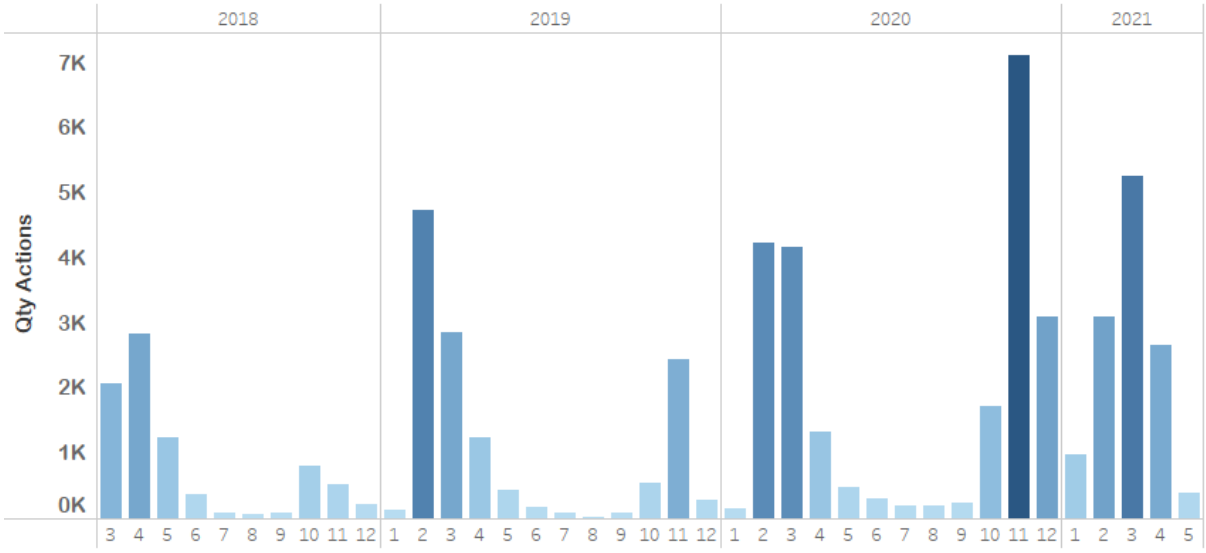

Figure 4: Logs Distribution.

On Table 5 also shows that the number of programs started has been increasing since 2018, and by the first months of 2021, it had already surpassed the 2019 figures, indicating that some numbers will be slightly higher than 2020. Naturally, program completion rates have followed enrollment rates, and as enrollments rise, so do completion rates, though the gap between the two is growing. Using the two years we have completed as an example, we have 1896 enrollments and 1653 program completions in 2019, giving us an 87 percent success rate, and 3559 enrollments and 2942 program completions in 2020, giving us an 83 percent success rate. The rate has been reduced slightly, but the values remain high. This dataset contains data from 20 different programs.

Table 5: Success Rate.

Action	2018	2019	2020	2021
Program Started	994	1896	3559	1873
Program Completed	903	1653	2942	1583
Success Rate	91%	87%	83%	85%

The second set of data has detailed information about the interactions that students have in each module of the course and the characteristics of each module. Lets' enunciate:

- *ID Fénix* - Fenix ID (from student);
- *ID Curso* - Program ID;
- *ID Versão Curso* - Program Version ID;
- *ID Módulo* - Module ID;
- *Tem Quiz?* - Has quiz?;
- *Número questões quiz* - Number of quiz questions;
- *Quiz necessário?* - Quiz required?;
- *Percentagem necessária para quiz* - Quiz required percentage;
- *Completou quiz?* - Completed quiz?;
- *Resultado quiz (%)* - Quiz result (%);
- *Tem vídeo?* - Has video?;
- *Vídeo necessário (%)* - Video required (%);
- *Progresso visualização (%)* - Visualization progress (%) ;
- *Tem inquérito?* - Has survey?;
- *Respondeu inquérito?* - Answered survey?;
- *Data Inscrição (Módulo)* - Enrollment Date (Module);
- *Data Finalização (Módulo)* - End Date (Module);
- *Data Inscrição (Curso)* - Enrollment date (Course);
- *Data Finalização (Curso)* - End Date (Course).

Table 6 allows us to examine the first five rows of the table. Due to the tables' size and number of columns, a transposed view was used to fit it into the document.

This table presents the information in detail at the module level and it contains information both on the details of each module as well as on the results obtained by the students in those modules.

There is also an increasing trend for enrollments throughout the years, as shown in Figure 5 in this data set, and data for 2021 is only available until mid-May.

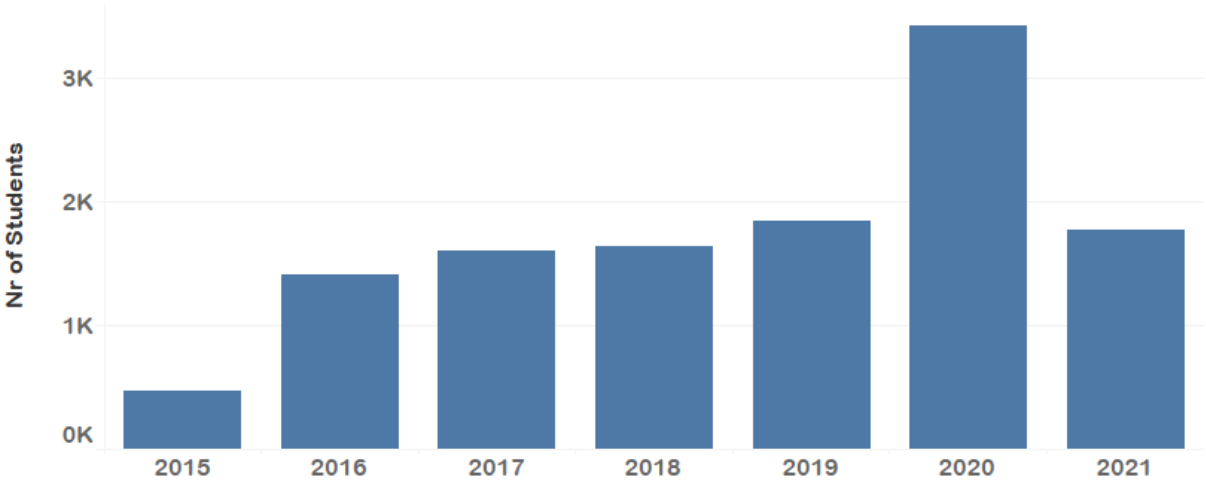

Figure 5: Enrollments by Year.

Table 6: Programs Dataset sample.

ID Fénix	2963528231177	2963528231177	2963528231177	2963528231177	2963528231177
ID Curso	4	4	4	4	4
ID Versão Curso	24795	24795	24795	24795	24795
ID Módulo	24781	24564	24566	24569	24571
Tem Quiz?	1	1	1	1	1
Número questões quiz	3	4	3	4	3
Quiz necessário?	1	1	1	1	1
Percentagem necessária para quiz	66	75	66	75	66
Completou quiz?	1	1	1	1	1
Resultado quiz (%)	100	100	100	100	100
Tem vídeo?	1	1	1	1	1
Vídeo necessário (%)	90	90	90	90	90
Progresso visualização (%)	0	0	0	0	0
Tem inquérito?	1	1	1	1	1
Respondeu inquérito?	0	0	0	0	0
Data Inscrição (Módulo)	2015-07-15	2015-07-15	2015-07-15	2015-07-15	2015-07-15
Data Finalização (Módulo)	2015-07-15	2015-07-15	2015-07-15	2015-07-15	2015-07-15
Data Inscrição (Curso)	2015-07-15	2015-07-15	2015-07-15	2015-07-15	2015-07-15
Data Finalização (Curso)	2015-07-15	2015-07-15	2015-07-15	2015-07-15	2015-07-15

The number of unique programs in this dataset differs from the previous one in that there are only 19 of them. Also, on this dataset, there is information since 2015, while on the previous the information was only beginning in 2018.

Figure 6 shows that almost every program has a different number of modules, averaging around 13 modules in a program.

As noticed on Figure 7, except for the year 2015, the information we obtain about the success rate in the programs dataset is identical to that obtained in the logs dataset, which is always around 90%. It is also noticed that, in 2016 and 2017, the majority of the programs was finished in the same day of the enrollment and on the following years

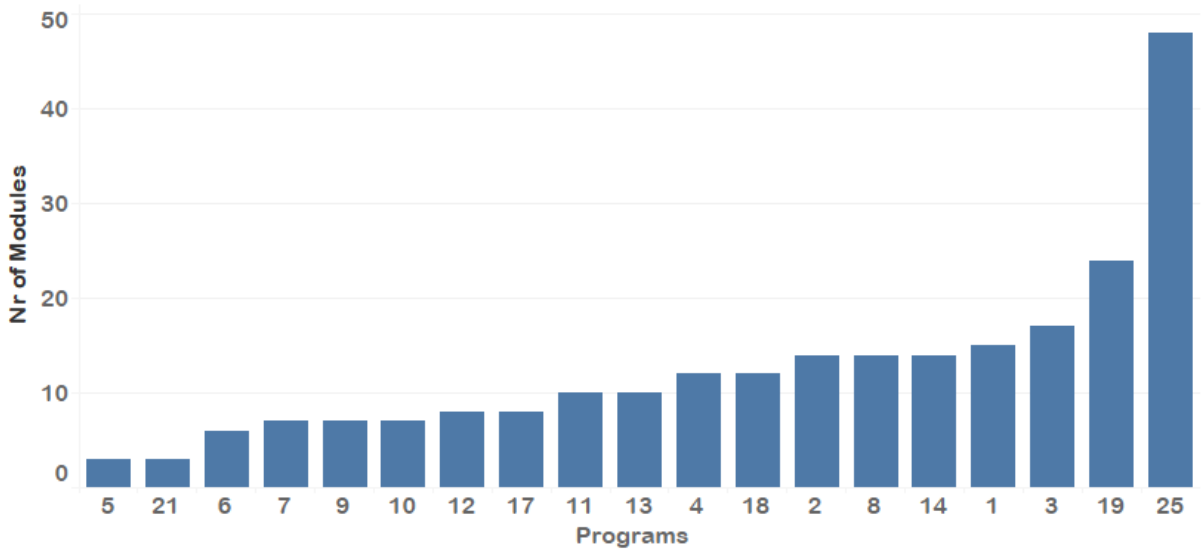

Figure 6: Modules by Program.

of 2018, 2019 and 2020 this rate is only around 40%. Finally, on 2021 is near 50%.

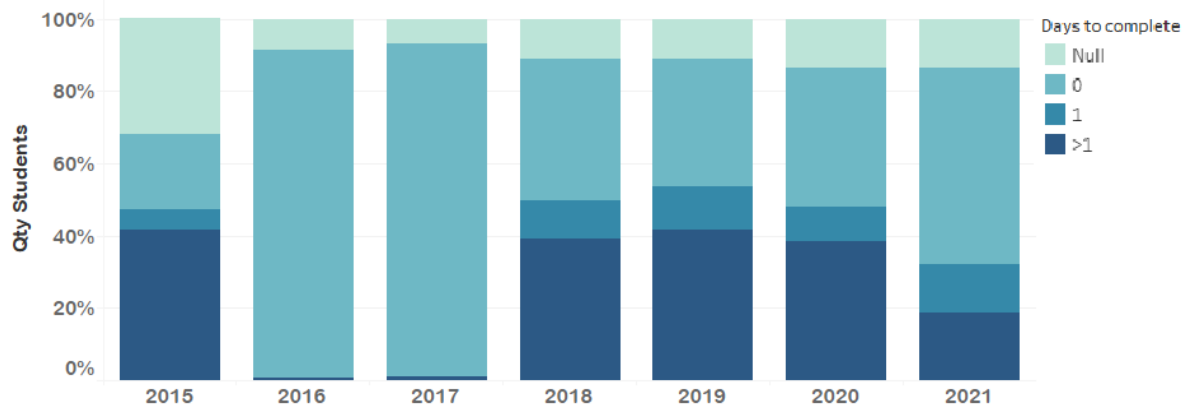

Figure 7: Success Rate on programs dataset.

The third dataset has information about the socio-demographic data of the students present in the previous sets. The characteristics are as follows:

- *ID Fénix* - ID Fénix;
- *Nacionalidade* - Nationality;
- *Segunda nacionalidade* - Second Nationality;
- *País nascimento* - Birth country;
- *Distrito nascimento* - Birth District;

Table 7: Info Dataset Sample.

Database field	Data sample
ID Fénix	2963528583739
Nacionalidade	Portugal
Segunda nacionalidade	
Pais nascimento	Portugal
Distrito nascimento	Guarda
País residência	Portugal
Distrito residência	Setúbal
Concelho residência	Setúbal
Condição profissional (aluno)	Desconhecido / Não tem
Sector profissional (aluno)	Desconhecido / Não tem
Profissão (mãe)	Oficial de Justiça
Condição profissional (mãe)	Trabalha por conta de outrem
Sector profissional (mãe)	Pessoal administrativo e similares
Habilitações literárias (mãe)	Ensino Médio
Profissão (pai)	Engenheiro Agrónomo
Condição profissional (pai)	Trabalha por conta de outrem
Sector profissional (pai)	Técnicos e profissionais de nível intermédio
Habilitações literárias (pai)	Ensino Pós-secundário - Curso de especialização Tecnológica

- *País residência* - Residence country;
- *Distrito residência* - Residence District;
- *Concelho residência* - Municipality Residence;
- *Condição profissional (aluno)* - Occupation (student);
- *Sector profissional (aluno)* - Profession sector (student);
- *Profissão (mãe)* - Profession (mother);
- *Condição profissional(mãe)* - Professional Status (mother);
- *Sector profissional (mãe)* - Profession sector (mother);
- *Habilitações literárias (mãe)* - Academic Qualifications (mother);
- *Profissão (pai)* - Profession (father);
- *Profissão (pai)* - Professional status (father);
- *Sector profissional (pai)* - Professional sector (father);
- *Habilitações literárias (pai)* - Academic Qualifications (father).

On Table 7 is shown a sample of the data in the dataset.

A few issues were discovered during the initial analysis of the dataset, which we will now describe:

- We can see that there are 373 records in the dataset that, despite not having a quiz, are required to take one. According to our research, everything is from course 5, and there are three modules (37446, 37450, 37453). We assume that the modules had parameterization errors or that they were finished before the requirement was defined;
- Although the quiz is not required to complete the module, there are 185 records in the dataset that have a minimum score of 75%. Because it is the same course and it is in the modules, this problem should be related to the one mentioned above;
- There are three modules with a 125 percent quiz result (course 7 with module 29032 and course 14 with module 37508) and one with a 150 percent result (course 7 with module 29032 and course 14 with module 37508). (course 4 with module 24584). It could be a problem with a modules' configuration (s);
- There are two modules whose completion date is earlier than the modules' enrollment date. (the students 2963528576698 in course 4, module 24781 and 2963528577420 in course 3, module 27996);
- The Course completion date is earlier than the registration date in 5 records. All of them are from the same class and the same student (Course 3 of student 2963528577994);
- There are more unemployed mothers in work status than there are unemployed mothers in occupation (438 Vs 277). The outcomes of the students' responses should most likely be the source of the problem.
- The column with the second nationality has few records with data on it in the socio-demographic dataset;
- Some dates are shown in the column with the birth district;
- There are some columns that are open-written and do not have typified text. To obtain information, more data cleaning is required.

The decision was made to delete the problematic records and proceed with the data treatment due to the low number of error occurrences recorded.

3.3 Additional calculated fields

Extra columns were added to the datasets to aid knowledge extraction.

In the dataset with information about the modules, two columns have been created. One to calculate the number of days between enrollment and module completion and another to calculate the number of days between enrollment and graduation in the program.

The column "action" was transformed into a column by type of action, marked with 1 when that action occurs and 0 when it does not, in order to get a treatment of the students' actions in the logs dataset.

3.4 Descriptive statistics

Some descriptive statistics were extracted from the datasets in order to gain a better understanding of the information received. For these, SPSS 27 was used.

3.4.1 Logs dataset

In the logs dataset there are 57049 records and with no missing values. These logs are spread over 20 different courses, 2 of which have two versions. Each log has one of the actions checked in Section 3.2. As shown in image 8, program 4 and 7 account for the majority of enrollments and conclusions, accounting for 50% of the dataset when combined.

As already shown on Table 5, 83% or more of the students in this dataset finish the course enrolled.

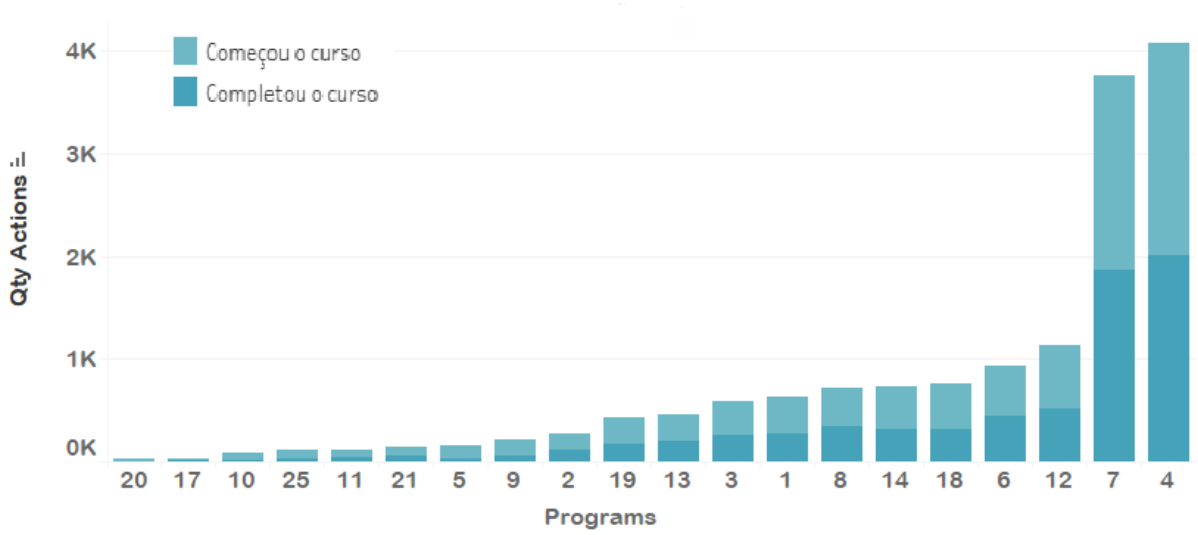

Figure 8: Actions per Program.

Surprisingly, we get more hits for the course completion certificate than we do for course completions. This is due to the fact that all accesses to the course completion certificate by students are counted. As an example of this, Table 8 shows the certificate access (*Acedeu ao Certificado*) number and the program completitions (*Completou o curso*)

Table 8: Certificate access sample.

Ação/Action	2018	2019	2020	2021
Acedeu ao certificado/Certificate access	1186	1727	3078	1775
Completou o curso/Program completed	903	1651	2939	1582
Difference	283	76	139	193

3.4.2 Programs dataset

There are 116027 observations in the program dataset. This figure is much higher than the previous one because it includes module-level information for each course.

In terms of the various characteristics of each module, we have quizzes in 95% of them, and mandatory quizzes in the same percentage of modules, with an average of 3.6 questions per quiz. As shown in Figure 9, only program 5, 10 and 21 have no questions (and quizzes). To pass the module, the quizzes must average a 73 percent score. Complete quizzes are available in 98 percent of the modules with a result of 93% on average.

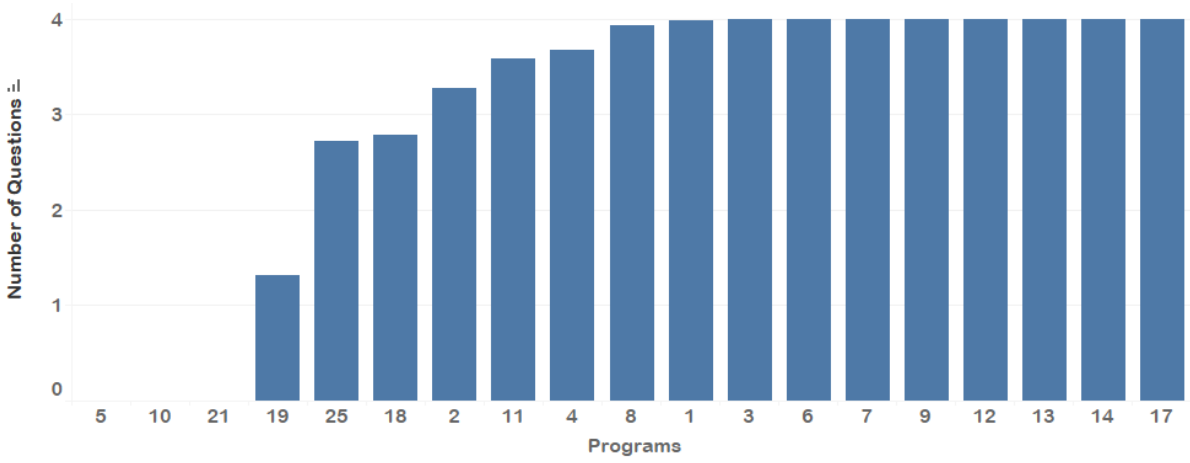

Figure 9: Questions per Quiz.

Every course has a video, and this video has a 70.8 percent of must view on average.

Surprisingly, each video is only seen 45 percent of the time. According to the explanation, the low value is due to an error in the software that reads the visualization progress.

There are surveys in 99% of the modules, and they are answered throughout 66% of the time.

Finally, students take an average of 0.8 days to finish a module and 10.4 days to complete a program.

Surprisingly or not, the modules with a quiz take less time to complete like the programs with a survey, as shown in image 10. This goes along with what is said in [11] where student engagement is essential to increase motivation and decrease the dropout rate.

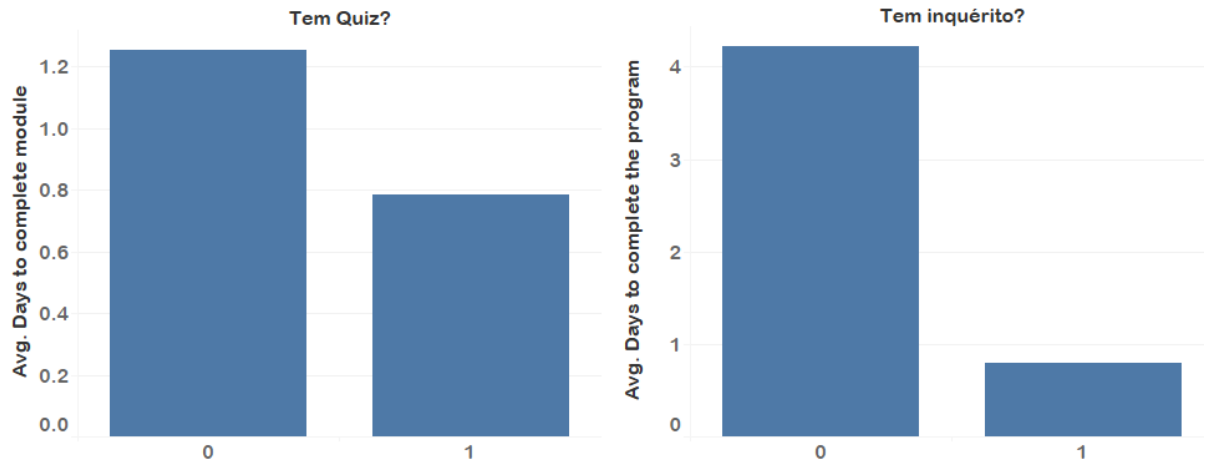

Figure 10: Days to Complete the Module or program.

3.4.3 Socio-demographic dataset

There are 6514 different student records in the dataset with socio-demographic data.

There are 16 blank records and 43 different nationalities in terms of nationality, with the majority being Portuguese (6059 records), followed by 60 Mozambicans, 56 Brazilians, and 51 Cape Verdeans.

The 6,238 blank registrations for the second nationality are the most significant aspect. We still have 41 Brazilians with a second nationality and 35 Angolans with a second nationality.

With 5790 records, Portugal is once again the most prominent birth country. Brazil comes in second with 91 records, followed by Mozambique with 84, Cape Verde with 64, Angola with 61, and Guinea-Bissau with 49. France, the first non-Portuguese-speaking country on the list, is ranked seventh. The number of null records is 14.

Lisbon has 3316 occurrences in the column of birth district, followed by Setubal with 600. We also have 334 records for Santarém and 302 records for Leiria. There are still

Table 9: Country of birth sample.

Portugal	Brazil	Mozambique	Cape Verde	Angola	Guinea-Bissau	France
5790	91	84	64	61	49	45

474 records that are blank.

There are 21 different countries on residence country column, with Portugal leading the way with 6386 records (98%). There are also 21 residents in France, 17 in Germany, 12 in Mozambique and Italy, and 11 in Brazil. There is only one empty record in this column.

There are 57 different records for the district of residence, with Lisbon having the most (58%), followed by Setubal, Santarém, and Leiria with 837, 351 and 318 records, respectively. The island of Madeira, with 147 records and 118 empty records, is also worth mentioning.

We have a greater dispersion of data in the county of residence column, with 246 different records. With 1163 records, Lisbon remains the most prominent. On this column there are 121 null records

Table 10: County of Residence sample.

Lisboa	Sintra	Cascais	Oeiras	Odivelas	Loures	Almada
1163	570	322	296	282	262	230
18%	9%	5%	5%	4%	4%	4%

There are 11 different types in the students' professional status column. In addition to the 87 null records, we highlight the 3144 who are students, the 2303 who have the condition as unknown, and the 671 who are Employees.

In the professional sector of the students, there are 12 different categories, with nearly 3/4 of the records being unknown / do not have and 15 percent having a different situation, leaving only 10% of the data set having some information.

We have 1346 different records in the mothers' occupation column and 1061 empty records. 531 teachers, 335 unemployed, and 209 housewives remain 8, 5 and 3 percent respectively.

We have the same 11 categories in the column for the professional status of the mothers' like what was found in the column for the students' professional status. In this case, the distribution is different, with 64% (4147) of the records indicating employment. 7% (438) are self-employed with employees, while 6% (402) are unemployed.

There are 12 different categories for the mothers' employment status, with 87 null records. There are 1378 people in the other situation out of the 12 categories, and 973 people who are not sure (21 and 15 percent respectively). There are also 901 (14%) as

Table 11: Mother Education.

Education	Count	%
Ensino Superior - Licenciatura (Pré-Bolonha)	1860	29%
Ensino Secundário - 12.º ano de escolaridade ou equivalente	1701	26%
Ensino Básico 3.º ciclo - 9.º ano de escolaridade (antigo 5ª ano liceal ou ensino técnico)	701	11%
Ensino Pós-graduado - Mestrado (pré-Bolonha)	392	6%
Ensino Superior - Bacharelato	331	5%
Ensino Básico 2.º ciclo - 6.º ano de escolaridade (antigo 2ª ano liceal ou ciclo preparatório)	298	5%
Ensino Básico 1.º ciclo - 4.º ano de escolaridade (antiga 4ª classe)	246	4%
Ensino superior - Licenciatura 1º ciclo (Bolonha)	231	4%
Desconhecido	154	2%
Ensino Pós-graduado - Doutoramento (pré-Bolonha)	109	2%
Ensino Médio	85	1%
Ensino Superior - Mestrado Integrado	78	1%
Sabe ler sem possuir o 4.º ano de escolaridade (antiga 4ª classe)	63	1%
Ensino Pós-secundário - Curso de especialização Tecnológica	59	1%
Ensino Pós-graduado - Mestrado 2º ciclo (Bolonha)	53	1%
Ensino Pós-graduado - Doutoramento 3º ciclo (Bolonha)	27	0%
Não sabe ler nem escrever	20	0%
Diploma de curso técnico superior profissional	19	0%
Total	6427	100%

administrative staff and 785 (12%) senior civil servants.

We have 18 categories for the educational backgrounds of the mothers of the students. The vast majority of records are from pre-Bologna degrees, which account for 29% of all occurrences, and 12th grade, which accounts for 26%. There are also 11% of records with a 9th grade level.

We have 1578 different records for the fathers' occupation, with the 1062 empty records having the most significant value (16%). Apart from the empty records, there are 282 for the businessman, 192 for the teacher, and 191 for the unemployed.

There are 11 different categories in the fathers' employment status column. The same as for students and mothers. Employees account for 54% of the occurrences (3059 records) in this case. 15% of fathers are self-employed with employees and 8% are self-employed without employees (985 and 506 records respectively). Note also that there are 87 empty records.

There are 12 different categories in the column for parents' employment status, and there are 87 empty records. However, we have 35% of the records distributed by other circumstances (22%) and unknown / have not (13%), which does not provide much insight. Despite this, there are 831 records for senior civil servants, 803 records for intermediate technicians, and 687 records for service and sales personnel.

Finally, there are 18 different categories in the column of parental educational attainment. The qualifications of the mothers are the same categories. There are 87 empty records in this case, and the distribution highlights 1706 records with a 12th grade (26%), 1303 records with a pre-Bologna diploma (20%), and 907 records with a 9th grade (14%).

To round out the data information, appendix A contains a list of the number of observations per result obtained, by the students, at the end. We averaged the characteristic values of each program module in the list and thus presented one observation for each student in a single iteration with the various programs. As it can be seen in the tables, almost every characteristic has a bigger number of observations with the success than with the ones marked as dropouts. The dataset has a lower number of observations as dropout when comparing to the number of observations with success result.

Chapter 4

Data Preparation and Modeling

This chapter explains how we are identifying students who are on the verge of dropping out. The themes mostly referred to MOOC-type programs, according to the literature reviewed. Because the characteristics of the Iscte - Instituto Universitário de Lisboa soft skills programs differ from those of MOOC programs and because the data obtained from transversal skills courses differs, there will need to be some adjustments in the data and algorithm selection.

The experiments will be divided in 3 scenarios. On Scenario A, the data will be randomly divided in a train and a test set. In order to be closer to a real case solution, on Scenario B, the train set will be the older years of the dataset and the newer years will be the test set. The Scenario C it will be like the previous scenarios,but will be an attempt to balance the targets in the dataset so that there are the same amount of success and dropout observations in the train dataset.

We will start by preparing the data, cleaning and selecting the characteristics in Section 4.1, then, on Section 4.2 go over modeling the data so the dataset is prepared to fit the models. On Section 4.3 are described the tests made with the different algorithms tested on the first scenario. On Section 4.4 It will be a different approach using the second scenario and close on Section 4.5 is a final approach using scenario C.

4.1 Data Preparation

This section discusses how to prepare the datasets so that they can be used with the models that have been tested. The data cleaning process will be described first, followed by the selection of features to use.

4.1.1 Data Cleaning

As previously stated, the only change made to the logs dataset was to create one column per feature of the type of actions recorded, with the column containing the type of action recorded in that row set to one and the rest set to zero.
Some data cleansing has been done on the program dataset:

- In the percentage required for the quiz, records that indicated a minimum percentage for the quiz were cleaned up using the column that indicated the quiz was not required. This was indicated to be a module parameterization error;
- When the modules had a result of more than 100%, it was set to 100%. This was yet another module parameterization error;
- There were only a few instances where the percentage of video viewed was greater than 100%. It had been set to 100%. The indication was that the software that reads the viewing progress had a configuration error;
- As previously mentioned, two columns were added to determine the number of days between enrollment and completion of modules and courses, thereby generating information from the dates.

Adjustments and cleaning were also made to the socio-demographic dataset:

- Four records containing dates were cleaned in the district of birth column;
- There was an entry with a number in the residency district column that was cleared;
- There was an entry in the county of residence column that only contained special characters that had to be cleaned up as well;
- There were some entries in the mothers' profession column that needed to be cleaned up, such as dots or dashes;
- There were some fields in the fathers' profession column where the same professions were written in a different way, so they were adjusted to be the same.

4.1.2 Feature Selection

The results obtained during the research studies are mostly about MOOC-type programs, and the types of data obtained differ slightly from those used in the researches. As a result, the characteristics chosen to the data models will need to be adjusted.

Contextual characteristics are important, as stated by [11], and also behavioral characteristics are important, according to [37]. Despite the fact that the data provided did not have the same exact characteristics, we can still use the ones we have that fit these

elements. Based on this, On Table 12 contains a summary of the features selected and the reasons why they are selected or excluded.
The module component was removed, leaving one line per program, to avoid overfitting with the repetition of the various socio-demographic features of the students and the program characteristics.

The models' target will be the time difference between the start and finish dates of the program. If the number of days exceeds a year (365 days) or is marked as incomplete, it will be considered a dropout and will be labeled as 0. Before a year has passed, the programs that have been completed will be marked as a success and labeled as 1. The variable is labeled like *Dias Curso*

4.2 Data Modeling

The difference between the day of enrolment and the day of competition will be used as the target variable because the goal of our models is to predict student dropout (or success). Because the student behavior will be classified using classification models, if the difference in dates is less than a year, the student will be classified as successful and the result will be a 1. Otherwise, it will be considered a dropout and a 0 will be assigned to it.

The data was divided into a set of 22 features, since the ID from the student was removed from the features, and a target before being used by the various algorithms. On the features dataset, we began by encoding the categorical variable using a method that creates a column for each different category and assigns 1 to the column of the category name on each row, while assigning 0 to all other categories. On the end of the encoding there was a dataset with 484 features Then, on all features, a scaler was used to ensure that they were all the same weight on the model. The *MinMaxScaler* method was used, in which each value was subtracted from that column's minimum value and divided by the difference between that column's maximum and minimum value. All values fit between 0 and 1 using this method. Finally, using a random split of 40% for testing, the features and target dataset was divided into a train and a test dataset.

As previously stated, the documentation found did not pertain to the same types of programs, but due to their proximity, we will use the models that produced the best results in this type of situation as a starting point. As mentioned in [11], good results were obtained with logistic regression classification, as well as good results with Supported Vector Machines (SVM) and Decision Trees in [22]. Random Forests are also used with good results in [4]. For these algorithms, the Python programming language was chosen, along with the scikit-learn library, which is one of the most widely used in machine

Table 12: Features Selected.

Feature	Selected	Motive
Logs		
ID do Fenix	No	It does not add anything to the data.
Id do Curso	No	It does not add anything to the data.
ID da Versão do Curso	No	It does not add anything to the data.
Ação	No	It does not add anything to the data.
Data	No	It does not add anything to the data.
Courses		
ID do Curso	Yes	Essential as an identifier
ID do Módulo	Yes	Essential as an identifier
Tem Quiz?	Yes	
Número de questões do quiz	Yes	
Quiz necessário?	Yes	
Percentagem necessária para o quiz	Yes	
Resultado do quiz (%)	Yes	
Tem vídeo?	Yes	
Vídeo necessário (%)	Yes	
Tem inquérito?	Yes	
Respondeu ao inquérito?	Yes	
ID do Fénix	No	Is used only as a link to the next dataset
ID da Versão do Curso	No	It does not add anything to the data.
Completou o quiz?	No	Does not add beyond the Resultado do quiz (%) column
Progresso de visualização (%)	No	It contains data with errors from the collect the data.
Data incrição (Módulo)	No	The number of days between enrolling and completing a module is used.
Data Finalização (Módulo)	No	The number of days between enrolling and completing a module is used.
Data incrição (Curso)	No	The number of days between enrolling and finishing a course is used.
Data Finalização (Curso)	No	The number of days between enrolling and finishing a course is used.
Dias_Módulo	No	To avoid overfitting
Info		
Nacionalidade	Yes	
País Nascimento	Yes	
Distrito Nascimento	Yes	
País Residencia	Yes	
Distrito Residencia	Yes	
Setor Profissional (Aluno)	Yes	
Condição Profissional (Mãe)	Yes	
Setor Profissional (Mãe)	Yes	
Habilitações Literárias (Mãe)	Yes	
Condição Profissional (Pai)	Yes	
Setor Profissional (Pai)	Yes	
Habilitações Literárias (Pai)	Yes	
ID do Fenix	No	Is used only as a link to the previous dataset
Segunda Nacionalidade	No	Information is in short supply
Concelho Residencia	No	There is far too much information that is stale and of poor quality.
Condição Profissional (Aluno)	No	There is far too much information that is stale and of poor quality.
Profissão (Mãe)	No	There is far too much information that is stale and of poor quality.
Profissão (Pai)	No	There is far too much information that is stale and of poor quality.

learning.
Some studies have used neural networks, such as [36, 37, 22, 13]. The results were also good, indicating that there would be some algorithms to explore. Python was used as the programming language in this algorithm, but we chose the Keras library because it usually produces better results with neural networks algorithms.

4.3 Comparing different models for predicting dropout (Scenario A)

One of the issues found in the dataset provided when preparing the data is that it is very unbalanced, with the majority of students having successfully completed the programs they enrolled in (close to 90%), leaving very few cases where the dropout occurred.

On the rest of this section we present the results obtained with each of the algorithms mention above, ending the section with a comparison of the results achieved.

4.3.1 Logistic Regression Classification

Although the logistic regression classifier is one of the most basic, the results obtained are already promising when predicting success, as shown in Table 13 where we can see that the student is classified as having completed the program with 96% of precision, but dropout is only predicted with 82% of precision. Despite this, the algorithm had a 96% accuracy rate.

Table 13: Logistic Regression Classification Report (Scenario A).

	Precision	Recall	F1-Score	support
Dropout	0.831	0.265	0.401	446
Success	0.961	0.997	0.979	8105
accuracy			0.959	8551
macro avg	0.896	0.631	0.690	8551
weighted avg	0.954	0.959	0.949	8551

The confusion matrix in Figure 14 shows that the predictions were more accurate on the real success cases, but the cases of dropout were predicted as success.

The best results were obtained with the "newton-cg" solver. The other solvers were not far but all in a lower level. The results were not significantly affected by any of the other parameters.

Table 14: Logistic Regression Confusion Matrix (Scenario A).

	Predicted dropout	Predicted success
True dropout	118	328
True success	24	8081

4.3.2 Support Vector Classification

We were able to improve both success and failure predictions using the Support Vector Classifier (SVC). The results of the classification report, as shown in Table 15, can be analyzed.

Table 15: SVC Classification Report (Scenario A).

	Precision	Recall	F1-Score	support
Dropout	0.793	0.291	0.426	446
Success	0.962	0.996	0.979	8105
accuracy			0.959	8551
macro avg	0.878	0.644	0.703	8551
weighted avg	0.953	0.959	0.950	8551

In this case, we were able to achieve a success precision rate of 96% and a dropout precision rate of 79%, with a median accuracy of 96%.

Table 16: SVC Confusion Matrix (Scenario A).

	Predicted dropout	Predicted success
True dropout	130	316
True success	34	8071

When comparing the results obtained with the previous algorithm, they are very similar. Figure 16 shows the confusion matrix.

The kernel is the only thing that gets results when tuning the classifiers' parameters. With the "linear" kernel, the best results were obtained.

4.3.3 Decision Trees Classification

We get better results with the decision trees algorithm than with Logistic Regression and SVC when classifying whether a student will complete the course or drop out. The

Table 17: Decision Tree Classification Report (Scenario A).

	Precision	Recall	F1-Score	support
Dropout	0.599	0.617	0.608	446
Success	0.979	0.977	0.978	8105
accuracy			0.958	8551
macro avg	0.789	0.797	0.793	8551
weighted avg	0.959	0.958	0.959	8551

Percision, Recall, F1-Score, and Accuracy values are always slightly higher, as shown in Table 17.

Although the confusion matrix has some values as false positives and false negatives, as shown in Figure 18, the value of dropouts obtained is higher than that obtained with SVC.

Table 18: Decision Trees Confusion Matrix (Scenario A).

	Predicted dropout	Predicted success
True dropout	275	171
True success	184	7921

The best results were obtained with the "entropy" criterion and the "random" splitter combined.

4.3.4 Random Forest Classification

Although the Random Forest algorithm is an evolution of Decision Trees, the results obtained in the classification of students who successfully completed the program are slightly higher, but the classifications of dropouts are lower. Table 19 illustrates this.

Table 19: Random Forest Classification Report (Scenario A).

	Precision	Recall	F1-Score	support
Dropout	0.862	0.422	0.566	446
Success	0.969	0.996	0.982	8105
accuracy			0.966	8551
macro avg	0.916	0.709	0.774	8551
weighted avg	0.963	0.966	0.961	8551

In Figure 20, we can see that the predictions of the programs' conclusion are very good, with a low number of false positives, but that the false negatives are higher in the case of dropout forecasts.

Table 20: Random Forest Confusion Matrix (Scenario A).

	Predicted dropout	Predicted success
True dropout	188	258
True success	30	8075

A maximum depth of 500 was used in the Random Forest, as well as a random state of 0, the "entropy" criterion, and 20000 as the number of estimators.

4.3.5 Neural Networks

There are different types of neural networks. The Artificial Neural Network (ANN), the Recurrent Neural Network (RNN) and the Convolution Neural Network (CNN) are the most common. The CNN are normally used with video and images files, while the RNN are mainly used for audio, time series and text data. So, if the problem we are handling is neither of these, an ANN is the solution to try.

The Keras library was used to implement the solution, as previously stated. Three dense layers were used. The first two are Relu-activated input layers. The final layer is a Sigmoid-activated output layer. A Categorical Cross Entropy Loss and an Adam optimizer were used to create the model. The model was fitted with 150 epochs, batch size of 10, and data spit with 40% used for testing.

Even with this model, the most difficult part was predicting which students would be dropouts. The F1-score of the dropout predictions is only 33.5%, as shown in Table 21.

Table 21: Neural Network Classification Report (Scenario A).

	Precision	Recall	F1-Score	support
Dropout	0.552	0.241	0.335	446
Success	0.958	0.989	0.973	8105
accuracy			0.948	8551
macro avg	0.755	0.615	0.654	8551
weighted avg	0.936	0.948	0.938	8551

This models' confusion matrix also shows that false positives and false negatives are both high, especially when compared to true negatives. Table 22 illustrates this.

This model also performs poorly in predicting these situations due to the lack of dropout examples in the dataset. On the other hand, when it comes to predicting success, the performance of this model, such as other models, is remarkable.

Table 22: Neural Network Confusion Matrix (Scenario A).

	Predicted dropout	Predicted success
True dropout	105	341
True success	96	8009

4.3.6 Summary of the Results

After running the different models in the Scenario A, a comparison must be made of the Performance, the Recall and F1-Score of each model.

The precision and recall of both success and dropout results were evaluated in order to establish the comparison. The precision findings reveal how many positive results were really ranked, whereas the recall results reveal how many positive outcomes should be ranked.The harmonic mean of precision and recall is shown on f1-score. This is reflected later in the results obtained in the confusion matrix.

As show in Figure 15, all models had a good performance when predicting the success. Both in terms of Precision and Recall, and, of course, in terms of the F1-Score. even so, the Neural Networks had always the worst result and the not as good result of the Decision Trees in recall is compensated by being the best result in precision.

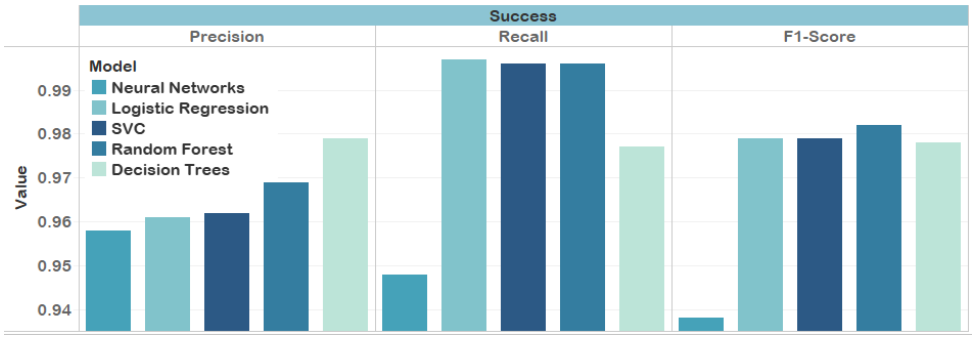

Figure 11: Models Success Performance Scenario A.

The problem is when analyzing the Dropouts and since is our major concern, as shown in Figure 12, in terms of precision, the Random Forest Classifier is the best model, followed by Logistic Regression and Support Vector Classifiers. This means that the Random Forest model has an 86.2% precision of assurance when classifying a student as a dropout.

The biggest problem is when analyzing the Recall of the different models. In this case, the best model is the Decision Trees and with a good margin.

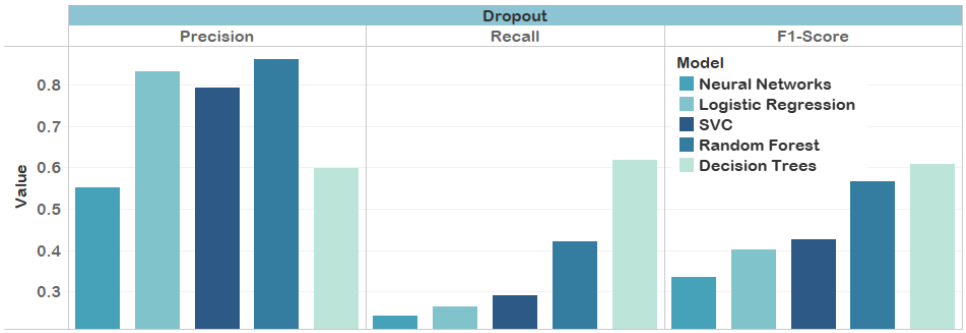

Figure 12: Models Dropout Performance Scenario A.

It has 61.7% and the second best is the Random Forest with 42.2%. This makes the Decision Trees the best model in terms of the sensitivity. The F1-Score was achieved by a combination of these results. Since this measure is the harmonic mean between the Precision and the Recall, the Decision Trees is the best model with 60.8%, followed by the Random Forest with 56.6%.

On Figure 13 is shown the correct predictions of each model, both predicting success and dropout.

On predicting success, all the models could correctly predict almost the totality of the cases. Once more is visible that the problem was to correctly predict the dropout. Nevertheless, the model with best results was the Decision Trees with 61.6% correct prediction. The second best was the Random Forest with 42.2% of correct predictions.

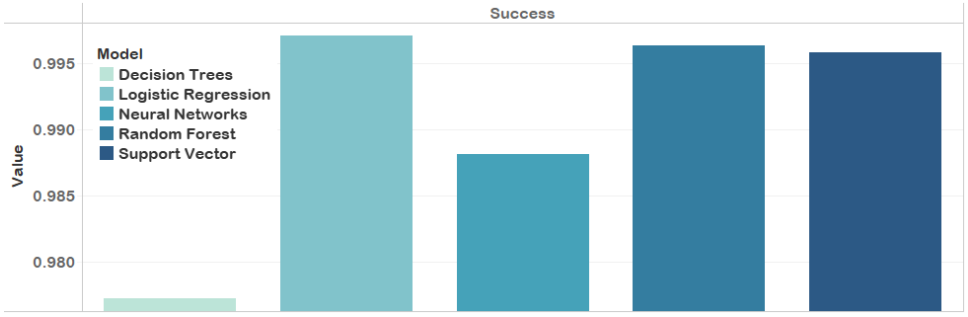

Figure 13: Success Correct Predictions Scenario A.

In terms of correct predictions of Dropout, as is shown in Figure 14, the result are not so good. The Decision Trees has the best result, but has only a 61.6% of correct predictions and the second best is the Random Forest with 42.1%.

When using the scenario A, the best model to use is the Decision Trees both by the

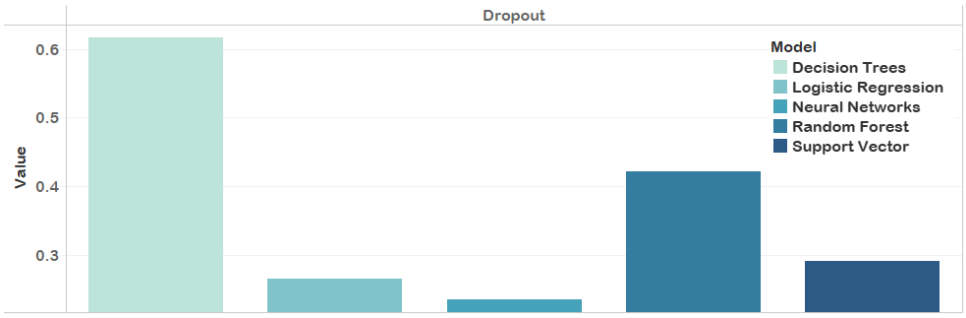

Figure 14: Dropout Correct Predictions Scenario A.

percentage of correct predictions and the overall performance stated by the F1-Score. Although using the Random Forest classification, if a student is classified as Dropout, it has a 86.2% of precision that is correct, so it is a good decision to also run these models and analyze the results.

Another way to address the problem is by excluding the students that are for sure a success student. If they have success in the program, they can not be a dropout. For this path, the best model is still the Random Forest with 96.9% of precision and 99.6% of Recall in predicting the success.

4.4 A more realistic approach (Scenario B)

We decided to take a more realistic approach using the same dataset and algorithms. We used the first years of data to train the various algorithms, leaving the more recent data to test the results.

On the rest of this section we present the results obtained with each of the same algorithms, ending the section with a comparison of the results achieved

4.4.1 Logistic Regression Classification

The results obtained with Logistic Regression Classification are low, especially on the dropout prediction. Like is shown in Table 23 the Dropout F1-Score is only 15.8%.

Table 23: Logistic Regression Classification Report (Scenario B).

	Precision	Recall	F1-Score	support
Dropout	0.800	0.088	0.158	684
Success	0.851	0.996	0.918	3592
accuracy			0.851	4276
macro avg	0.826	0.542	0.538	4276
weighted avg	0.843	0.851	0.796	4276

On Table 24 there is a confusion matrix where is shown that the Dropout correctly predicted is very low where only 60 cases were predicted in a total of 684 real cases.

Table 24: Logistic Regression Confusion Matrix (Scenario B).

	Predicted dropout	Predicted success
True dropout	60	624
True success	15	3577

Once more, the only parameter that truly had some different results was the solver, but in this scenario, both newton-cg and saga solvers had similar results.

4.4.2 Support Vector Classification

With the Support Vector Machine Classification once more, when predicting the dropout, the performance is low. But in this case they were lower than the Logistic Regression. The accuracy obtained was 84.8% as is shown on Table 25, but the Recall value of the dropout is only 6.9%

Table 25: Support Vector Classification Report (Scenario B).

	Precision	Recall	F1-Score	support
Dropout	0.810	0.069	0.127	684
Success	0.849	0.997	0.917	3592
accuracy			0.848	4276
macro avg	0.830	0.533	0.522	4276
weighted avg	0.843	0.848	0.791	4276

As shown on Table 26, the correctly predicted dropout results were only 47 of 684 cases. Even lower than the Logistic Regression model.

Table 26: Support Vector Confusion Matrix (Scenario B).

	Predicted dropout	Predicted success
True dropout	47	637
True success	11	3581

Like in scenario A, the best kernel was the linear kernel. the other parameters for these model produced similar results to the default values.

4.4.3 Decision Trees Classification

The Decision Trees model results achieved are shown on Table 27. The accuracy is 85.6% which is the best so far with this scenario, but the dropout predictions is were the major

enhancement is, where with a low precision value is balanced with a good recall value, leading to 57.2% on the F1-Score.

Table 27: Decision Trees Classification Report (Scenario B).

	Precision	Recall	F1-Score	support
Dropout	0.545	0.601	0.572	684
Success	0.922	0.905	0.913	3592
accuracy			0.856	4276
macro avg	0.734	0.753	0.743	4276
weighted avg	0.862	0.856	0.859	4276

The confusion matrix shown in table 28 has a big improvement on predicting the dropout where 411 out of 684 results were correctly predicted. on the success predictions, this model performs worst than the previous ones.

Table 28: Decision Trees Confusion Matrix (Scenario B).

	Predicted dropout	Predicted success
True dropout	411	273
True success	343	3249

On this scenario, and as we are trying to predict the dropout, the best results were achieved with the "entropy" criterion and the "random" splitter, like in Scenario A, but to predict success, the "gini" criterion preformed better.

4.4.4 Random Forest Classification

The accuracy of Random Forest Classification is bigger than Decision Trees Classification. This is greatly achieved by the great precision in prediction on predicting success, but also predicting dropout. On the other hand, the Recall of the dropout predictions is very low. This turns out in a F1-Score of 24.9% for the dropout predictions which leads to a less accurate predictions on dropout.

On the confusion matrix shown in Table 30 the result that is positively better is the amount of successfully predicted success cases. On opposite result is the few cases that were successfully predicted as dropout.

The parameters used to these predictions were an "entropy" criterion, a random state of 0 and a maximum depth of 500, just like in Scenario A. The only parameter different were the number of estimators that were 10000 which is half of the used in Scenario A.

Table 29: Random Forest Classification Report (Scenario B).

	Precision	Recall	F1-Score	support
Dropout	0.942	0.143	0.249	684
Success	0.860	0.998	0.924	3592
accuracy			0.862	4276
macro avg	0.901	0.571	0.596	4276
weighted avg	0.873	0.862	0.816	4276

Table 30: Random Forest Confusion Matrix (Scenario B).

	Predicted dropout	Predicted success
True dropout	98	586
True success	6	3586

4.4.5 Neural Networks

The classification by Neural Networks, as shown in Table 31, achieved a 84.9% of accuracy which is a great result, but looking in more detail, the results obtained in predicting dropout were not so great with a precision of 61.2% and a Recall of 7.6% which ends in a result of 13.5% in F1-Score.

Table 31: Neural Network Classification Report (Scenario B).

	Precision	Recall	F1-Score	support
Dropout	0.612	0.076	0.135	684
Success	0.849	0.991	0.915	3592
accuracy			0.844	4276
macro avg	0.730	0.533	0.525	4276
weighted avg	0.811	0.844	0.790	4276

As shown in Table 32, the confusion matrix indicates that only 52 out of 684 cases were correctly predicted as dropout. On the success predictions, the results were much better with only 33 cases predicted incorrectly in a total of 3592.

For this scenario the same type of Neural Network was used as in Scenario A. The first two layers are input layers and are Relu-activated. The final layer is a Sigmoid-activated output layer. A Categorical Cross Entropy Loss and an Adam optimizer were used to create the model. The only change was the test value used in Scenario B.

4.4.6 Summary of the Results

Like in the previous scenario, a comparison of the performance has to be made between the different models. On Figure 15 is shown a chart with the measures of the success of each tested model.

Table 32: Neural Network Confusion Matrix (Scenario B).

	Predicted dropout	Predicted success
True dropout	52	632
True success	33	3559

Predicting the success achieved good results, and once more the model with best precision was the Decision Trees and followed by the Random Forest. In terms of Recall, all the models were much alike with the exception of the Decision Trees that was lower than the others. This result lead to a F1-Score that was almost the same in all models, with the Neural Network being the one which performed slightly better than the other models.

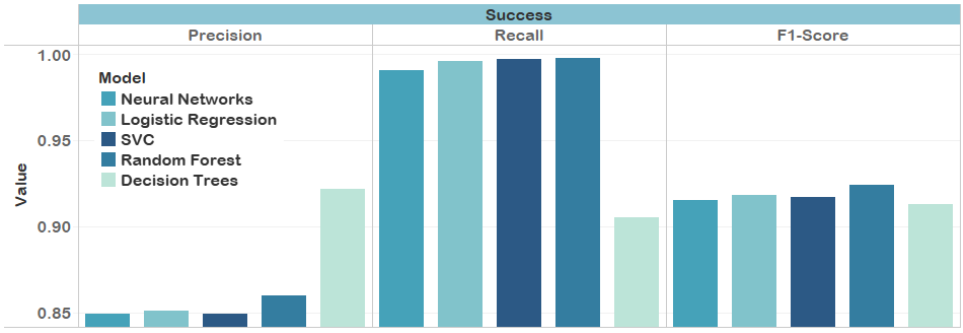

Figure 15: Models Success Performance Scenario B.

In term of measuring the dropout, as show in Figure 16 the model with better precision was the Random Forest, although is notable that, on this scenario, the Support Vector had a precision much closer to the Logistic Regression Classification and the model with lower performance was the Decision Trees.

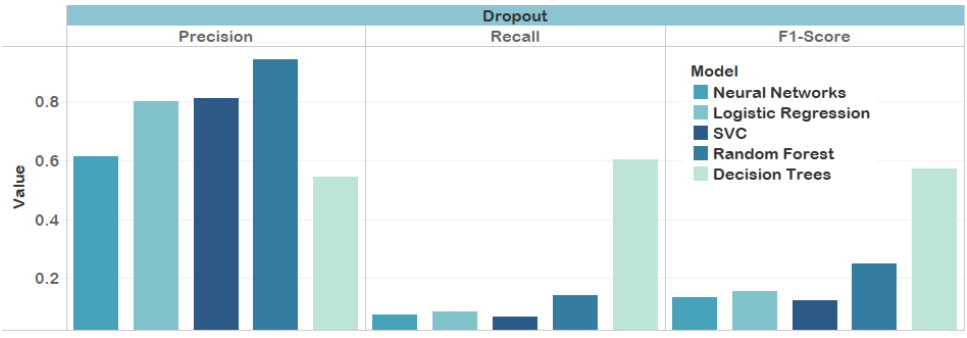

Figure 16: Models Dropout Performance Scenario B.

On the other hand, measuring the recall, all the models performed poorly with the decision trees being the only exception, having results of 60.1%. Still not great, but better than the others.

The F1-Score results, since all the models had very poor performance in recall with the exception of the Decision Trees that had a result near the precision, shows that the better model was the Decision Trees.

Like on the previous scenario, a comparison of correct predictions between the different models help us analyzing the performance of the models.

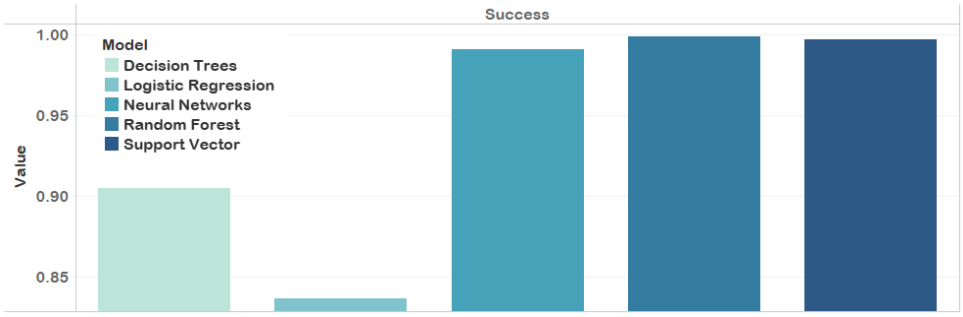

Figure 17: Success Correct Predictions Scenario B.

Scenario B Still had good results when predicting the success, but not in all similar. As show in Figure 17, the Random Forest had 99.8% and Support Vector models had and 99.6% which is almost every case. The other 3 models were not so high on predictions but still the Logistic Regression was the worst with 83.7%.

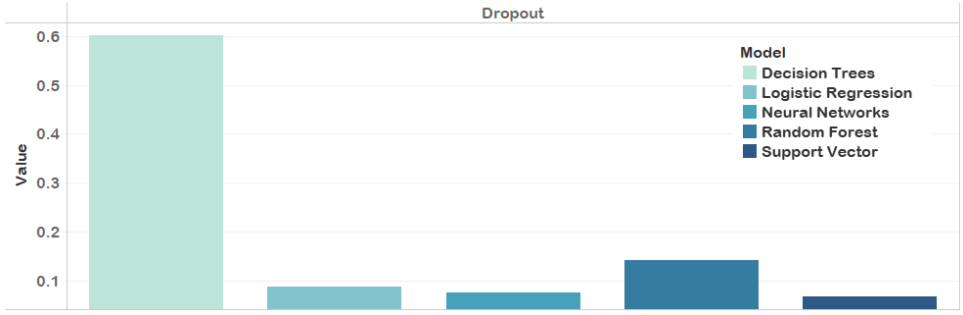

Figure 18: Dropout Correct Predictions Scenario B.

Once more the problem was predicting the dropout of the students. As shown in Figure 18, on Scenario B, the Decision Trees Classification achieved 60.1% of correct predictions. The second best model was again the Random Forest but on this scenario achieved only 14.3% of correct predictions. The remaining 3 models could not achieve 10% of correct answers.

Since the purpose of the research is finding the best model to predict the dropout, we find that, once more, the Decision Trees had the best performance of the models used.

It had the larger percentage of correct answers when predicting the dropout and still achieved over 90% when predicting the success.

The Random Forest had a great precision on predicting the dropout, but since it also had a very low performance in recall when predicting the dropout, the use of this model to predict dropout can lead to unsafe results since it has low sensitivity.

4.5 Dealing With the Unbalanced Data (Scenario C)

Since one of the major problems was that the dataset has much more cases of student success than dropouts, we decided to use one more library from Scikit-Learn to help us balancing the problem. The library is called SMOTE and works by oversampling the dataset on the minority class and therefore adds representation on that class without adding information to the model.

After balancing the dataset used on Scenario B, we decided to run once more the Decision Trees and the Random Forest models. As expected, both of them had significant improvements.

4.5.1 Decision Trees

With the dataset balanced, the success predictions with this model were not so great. It was expected, since the classes were balanced. Even so, the lost of performance on success predictions was minor and the gain on dropouts was significant. As shown on Table 33, the precision was now 60.9% and the recall was 56.7% while predicting dropout and 91.9% of precision and 93.1 while predicting success,

Table 33: Decision Trees Classification Report (Scenario C - Balanced).

	Precision	Recall	F1-Score	support
Dropout	0.609	0.567	0.587	684
Success	0.919	0.931	0.925	3592
accuracy			0.873	4276
macro avg	0.764	0.749	0.756	4276
weighted avg	0.869	0.873	0.871	4276

On Table 34 shows 388 correctly predictions of dropout out of 684 cases which is a 56.7% of accuracy in predicting the dropout and as 3343 correct predictions of success out of 3592 which is 93%.

Table 34: Decision Trees Confusion Matrix (Scenario C - Balanced).

	Predicted dropout	Predicted success
True dropout	388	296
True success	249	3343

Using the model balanced with the SMOTE library shows significant improvements in predicting the dropout and since the success predictions were only slight lower, it prove to be a good model.

On Figure 19 is a chart of the learning process of the model. In this case is shown that the validation was never too close to the training score and the model had a training time almost linear to the training data size. Is also shown that the training time is low.

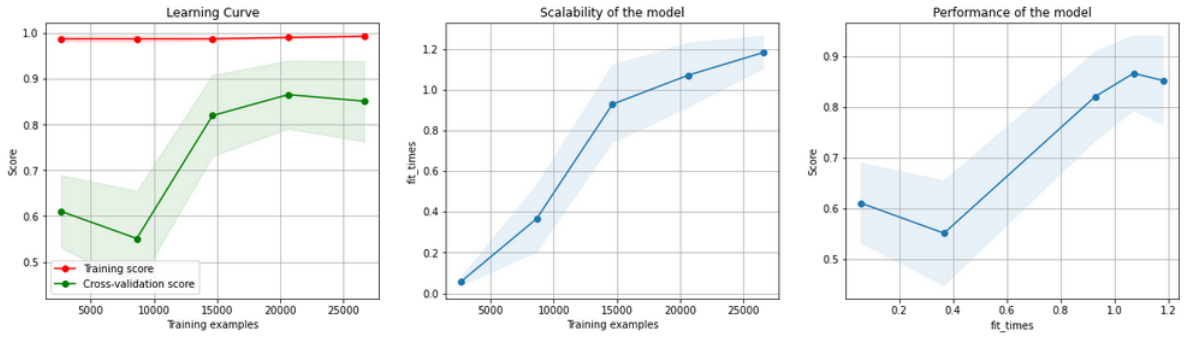

Figure 19: Decision Trees Training Chart

4.5.2 Random Forest

As the Random Forest Model was the second best and it had a close performance to the Decision Trees, we decided to also had a run with the balanced dataset. As shown in Table 35, once more the precision on predicting dropout was very high and the problem still remains in the recall. Also like the Decision Trees, the results were close but lower when comparing the Scenario B. The Dropout had 91.3% of precision as 13.9% of recall and the success had a precision of 85.9% and a recall of 99.7%.

As shown in Table 36, the correct predictions in dropout were 2214 correct predictions of 3592 which give us 61.6% which is also a good result but over 10% lower than the achieved with the Decision Trees model.

It is also shown in Table 36 that the correct predictions of success were 99.6%. This model still is a good option mainly for predicting success.

Table 35: Random Forest Classification Report (Scenario C - Balanced).

	Precision	Recall	F1-Score	support
Dropout	0.913	0.139	0.241	684
Success	0.859	0.997	0.923	3592
accuracy			0.860	4276
macro avg	0.886	0.568	0.582	4276
weighted avg	0.868	0.860	0.814	4276

Table 36: Random Forest Confusion Matrix (Scenario C - Balanced).

	Predicted dropout	Predicted success
True dropout	95	589
True success	9	3583

On this model it was also analyzed the learning curve like is shown in Figures 20. It is shown that the Cross Validation Score it is much more close to the leaning score and in terms of scalability is even more linear than the Decision Trees. One set back is the learning time spent is way more bigger.

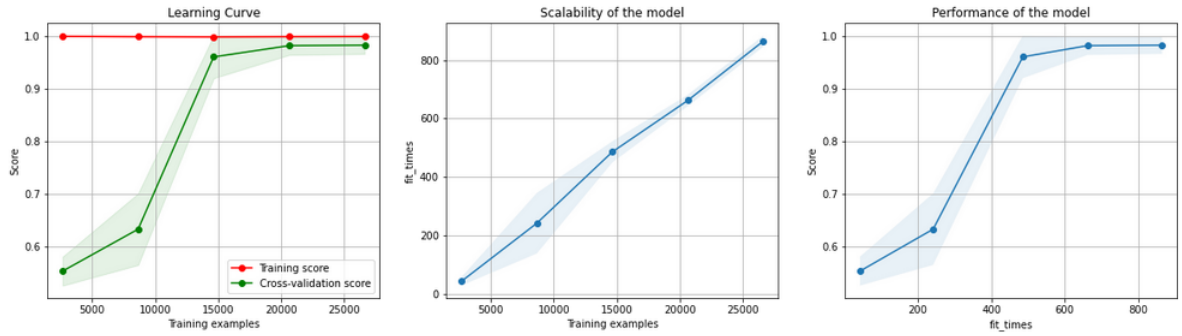

Figure 20: Random Forest Training Chart

4.5.3 Summary of the Results

After experimenting with the several algorithms and on Scenarios A, B and C, the Decision Trees and Random Forest classifications consistently produced the best results. On Figure 21 is shown the success behavior of this 2 models during the 3 scenarios used.

In terms of predicting success, the Random Forest had a similar performance on correct predictions in all scenarios mainly due to a recall that was almost equal in all scenarios.

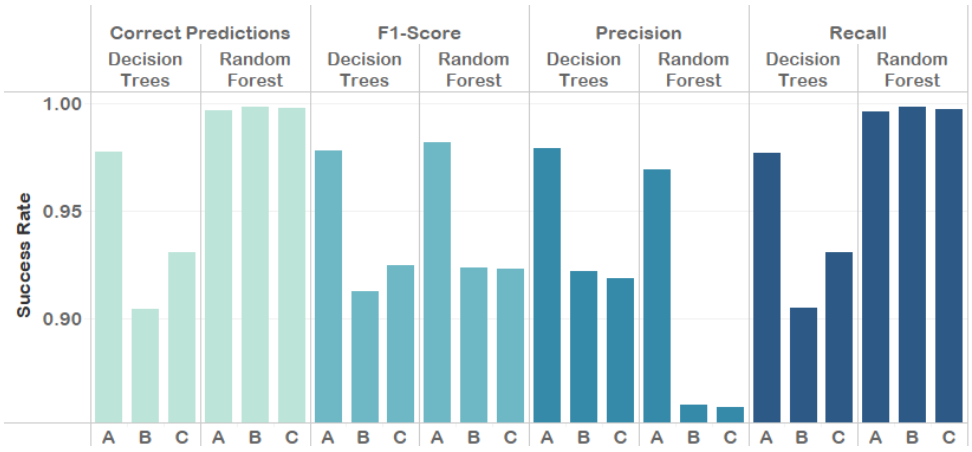

Figure 21: Results Evolution on Success

Both Precision and F1-Score measures decreased the results throughout the Scenarios evolution.

Focusing on predicting the dropout, it is shown on Figure 22, all the measures suffer a loss of performance from Scenario A to Scenario B, and then from B to C. The only exception is the precision of the Random Forest model that increased on Scenario B, but suffer a minor loss on Scenario C.

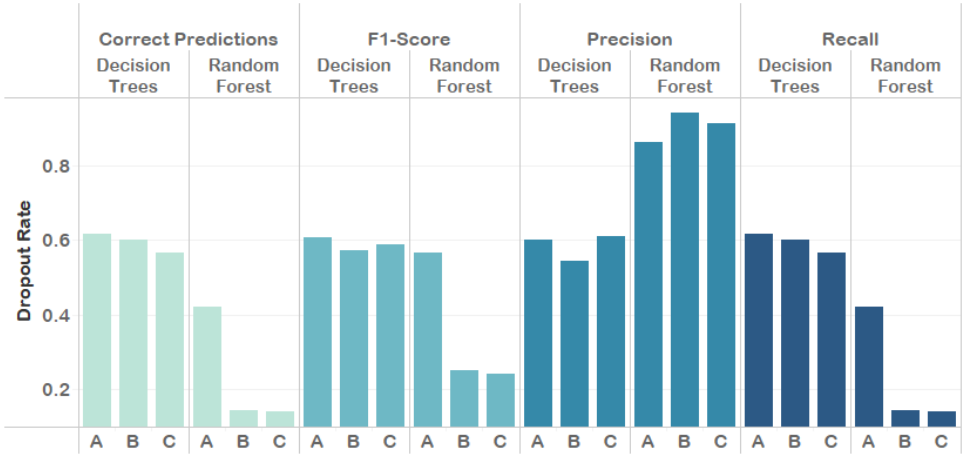

Figure 22: Results Evolution on Dropout

As shown on this charts the use of the SMOTE Library did not produced any better result when comparing to Scenario B.

Chapter 5

Conclusions

The high number of dropouts is one of the most serious issues for online and distance learning programs. A major cause might be the lack of interaction among students and between the student and the teacher, as well as the students' social and academic backgrounds.

The main goal of this study was to identify students who were at risk of dropping out early on. Data from Iscte - Instituto Universitário de Lisboa soft skills courses was provided to assist us in this research, and as a result of this work, we attempted to answer the questions presendted in section 1.3

Concerning the first research question, "What are the most effective characteristics for predicting student dropout in online education?", both from statistic metrics and form the decision trees library, we can see that the following features have a greater impact on our target variable (success or dropout):

- Video required (%) - *Vídeo necessário (%)*
- Quiz result (%) - *Resultado quiz (%)*
- Answered survey? - *Respondeu inquérito?*
- Profession sector (student) - *Sector profissional (aluno)*
- Professional Status (mother) - *Condição profissional (mãe)*
- Profession sector (mother) *Sector profissional (mãe)*
- Academic Qualifications (mother) - *Habilitações literárias (mãe)*
- Professional status (father) - *Condição profissional (pai)*
- Professional sector (father) - *Sector profissional (pai)*

- Academic Qualifications (father) - *Habilitações literárias (pai)*

As previously demonstrated, all of the models performed well when it came to predicting success cases. This in line with what we expected since the proportion of success is much greater than the dropout. When comparing the different models, the Decision Trees Classification had the best performance in terms of predicting dropout, but the Random Forest model provide us with a great precision predicting student dropout. The problem with the Random Forest model was the recall value in dropout that was too low, so the number of correct predictions in dropout was low, although the ones predicted were with great assurance.

One of the advantages of the Decision Trees model over the Random Forest is the computing expense of each model. To achieve similar results, the Decision Trees performed much faster than the Random Forest. This generates new results more quickly.

Because predicting student dropouts, which is the focus of the study, has proven to be difficult, and because the results for predicting success have not been very different, we chose the Decision Tree algorithm as a response to our second research question, "Which algorithm provides us with better results in predicting student dropout in online education?". The results in predicting success are very similar to all models, but predicting dropout is the best model and it is much faster than the Random Forest Model that was the model most similar in terms of results.

To answer the 3rd research question, "How can we develop an early warning system for teachers regarding student dropout in online education?", since the model selected for obtaining is Decision Trees, we chose to draw up a complete map of the entire model during the various decision steps. Since the complete model is too extensive, we put it in Appendix B. In Figure 23 is a sample of the first 3 decision levels, where we can see that the first step is whether the respondent answered the questionnaire. If the value is greater than 0.49, the next step is to know the ID of the program in which the student enrolled. If it is lower than 0, 49 and therefore negative, it is necessary to know the quiz result.

We can then make a map with the remaining steps that are found in the Tree Map results, but due to the length of the Tree Map we will not describe all the steps. For the same reason a picture with the Tree Map was not provided in Appendix B, because due to the size the letters were too small, but it was provided in text form.

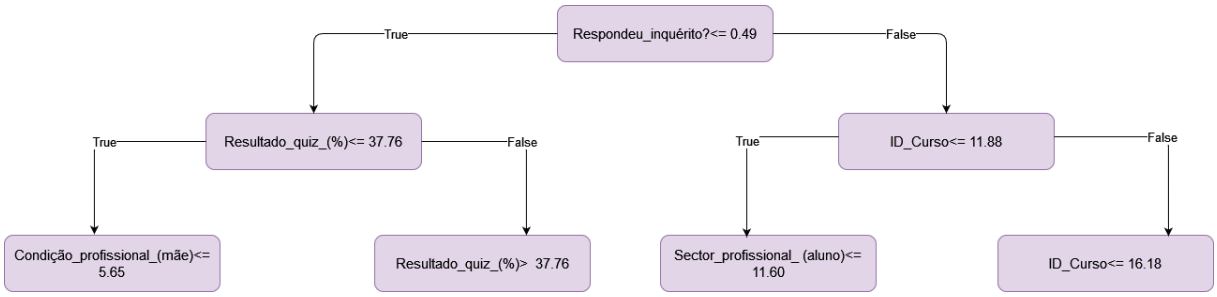

Figure 23: Decision Trees Result Sample

5.1 Limitations

One of the issues was that, while the programs were online, they differed from the MOOC (Massive Open Online Course) programs found in the study due to their short course length. To solve this, modifications had to be made so that searches could be based on similarities between online programs in a more general way.

Similarly, the data we were given was not identical to that found in the studies, so generalizing the concepts discovered made even more sense.

As previously stated, the data provided was not only of different quality than that of the studies conducted, but it was also over-represented in terms of success cases and contained a large number of drop-outs. Because of this, the success predictions performed well in all of the models used, while the dropout predictions performed poorly. To address this problem, the target classes were balanced using a Scikit-Learn library called SMOTE, and the results showed a slight loss of performance in predicting success, but a significant improvement in predicting dropout. This proved to be a good solution because the main goal of this study was to predict dropout.

When modeling the different scenarios with the Neural Networks it was found that this model is a too powerful tool to use with so little amount of data and therefore it was a lot of time consuming to produce so little results. This was a model to test once again and with better tuning, but when more data is created and added to the current dataset.

5.2 Future Work

After analyzing the results achieved, it cleared that the data provided was not enough to the problem suggested. It was cleared that the models would require more fine tuning so that would not be necessary the use of libraries to balance both classes. It was one of the biggest concern during the research.

When provided sufficient data both with dropout and success, the new data should

be modeled once more so that a warning system might be built that would send out early signals about students who were on the verge of dropping out.

Bibliography

[1] Carlos Alario-Hoyos et al. "Understanding learners' motivation and learning strategies in MOOCs". In: *The International Review of Research in Open and Distributed Learning* 18.3 (2017).

[2] Thushari Atapattu and Katrina Falkner. "Impact of lecturer's discourse for students' video engagement: Video learning analytics case study of moocs". In: *Journal of Learning Analytics* 5.3 (2018), pp. 182–197.

[3] William Bloemer et al. "Digging Deeper into the Data: The Role of Gateway Courses in Online Student Retention." In: *Online Learning* 22.4 (2018), pp. 109–127.

[4] Inma Borrella, Sergio Caballero-Caballero, and Eva Ponce-Cueto. "Predict and intervene: Addressing the dropout problem in a MOOC-based program". In: *Proceedings of the Sixth (2019) ACM Conference on Learning@ Scale*. 2019, pp. 1–9.

[5] Aras Bozkurt and Yavuz Akbulut. "Dropout patterns and cultural context in online networked learning spaces". In: *Open Praxis* 11.1 (2019), pp. 41–54.

[6] Hongyu Pei Breivold, Ivica Crnkovic, and Magnus Larsson. "A systematic review of software architecture evolution research". In: *Information and Software Technology* 54.1 (2012), pp. 16–40.

[7] Nergiz Ercil Cagiltay, Kursat Cagiltay, and Berkan Celik. "An Analysis of Course Characteristics, Learner Characteristics, and Certification Rates in MITx MOOCs". In: *International Review of Research in Open and Distributed Learning* 21.3 (2020), pp. 121–139.

[8] Chen Chen et al. "The impact of student misconceptions on student persistence in a MOOC". In: *Journal of Research in Science Teaching* 57.6 (2020), pp. 879–910.

[9] Zexuan Chen, Jianli Jiao, and Kexin Hu. "Formative Assessment as an Online Instruction Intervention: Student Engagement, Outcomes, and Perceptions". In: *International Journal of Distance Education Technologies (IJDET)* 19.1 (2021), pp. 1–16.

[10] Carol S Gering et al. "Strengths-based analysis of student success in online courses." In: *Online Learning* 22.3 (2018), pp. 55–85.

[11] Brenda Edith Guajardo Leal, Valenzuela GonzĆ, et al. “Student Engagement as a Predictor of xMOOC Completion: An Analysis from Five Courses on Energy Sustainability.” In: *Online Learning* 23.2 (2019), pp. 105–123.

[12] Liu Haiyang et al. “A time series classification method for behaviour-based dropout prediction”. In: *2018 IEEE 18th international conference on advanced learning technologies (ICALT)*. IEEE. 2018, pp. 191–195.

[13] Yanbai He et al. “Online At-Risk Student Identification Using RNN-GRU Joint Neural Networks”. In: *Information* 11.10 (2020), p. 474.

[14] Maartje Henderikx, Karel Kreijns, and Marco Kalz. “A classification of barriers that influence intention achievement in MOOCs”. In: *European Conference on Technology Enhanced Learning*. Springer. 2018, pp. 3–15.

[15] Maartje Henderikx et al. “Factors influencing the pursuit of personal learning goals in MOOCs”. In: *Distance Education* 40.2 (2019), pp. 187–204.

[16] Maartje A Henderikx, Karel Kreijns, and Marco Kalz. “Refining success and dropout in massive open online courses based on the intention–behavior gap”. In: *Distance Education* 38.3 (2017), pp. 353–368.

[17] Dan Yngve Jacobsen. “Dropping out or dropping in? A connectivist approach to understanding participants’ strategies in an e-learning MOOC pilot”. In: *Technology, Knowledge and Learning* 24.1 (2019), pp. 1–21.

[18] Giovani Lemos De CARVALHO JUNIOR et al. “Comparative study SPOC VS. MOOC for socio-technical contents from usability and user satisfaction”. In: *Turkish Online Journal Of Distance Education* 20.2 (2019), pp. 4–20.

[19] Roland Klemke, Maka Eradze, and Alessandra Antonaci. “The flipped MOOC: using gamification and learning analytics in MOOC design—a conceptual approach”. In: *Education Sciences* 8.1 (2018), p. 25.

[20] Roland Klemke et al. “Designing and Implementing Gamification: GaDeP, Gamifire, and applied Case Studies”. In: *The International Journal of Serious Games* 8.3 (2020).

[21] Daeyeoul Lee, Sunnie Lee Watson, and William R Watson. “The Influence of Successful MOOC Learners’ Self-Regulated Learning Strategies, Self-Efficacy, and Task Value on Their Perceived Effectiveness of a Massive Open Online Course”. In: *International Review of Research in Open and Distributed Learning* 21.3 (2020), pp. 81–98.

[22] Youssef Mourdi et al. “A machine learning based approach to enhance MOOC users’ classification”. In: *Turkish Online Journal of Distance Education* 21.2 (2020), pp. 47–68.

[23] Olga Pilli, Wilfried Admiraal, and Aysegul Salli. “MOOCs: Innovation or stagnation?” In: *Turkish Online Journal of Distance Education* 19.3 (2018), pp. 169–181.

[24] Eyal Rabin et al. “Identifying Learning Activity Sequences that Are Associated with High Intention-Fulfillment in MOOCs”. In: *European Conference on Technology Enhanced Learning*. Springer. 2019, pp. 224–235.

[25] Luis M Romero-Rodriguez, Maria Soledad Ramirez-Montoya, and Jaime Ricardo Valenzuela González. “Gamification in MOOCs: Engagement application test in energy sustainability courses”. In: *IEEE Access* 7 (2019), pp. 32093–32101.

[26] S Ros et al. “Analyzing students’ self-perception of success and learning effectiveness using gamification in an online cybersecurity course”. In: *IEEE Access* 8 (2020), pp. 97718–97728.

[27] Léon JM Rothkrantz. “New Didactic Models for MOOCs.” In: *CSEDU (1)*. 2017, pp. 505–512.

[28] Natalia Stathakarou et al. “MOOC learners’ engagement with two variants of virtual patients: A randomised trial”. In: *Education Sciences* 8.2 (2018), p. 44.

[29] Otgontsetseg Sukhbaatar, Tsuyoshi Usagawa, and Lodoiravsal Choimaa. “An artificial neural network based early prediction of failure-prone students in blended learning course”. In: *International Journal of Emerging Technologies in Learning (iJET)* 14.19 (2019), pp. 77–92.

[30] Ayse Saliha Sunar et al. “How learners’ interactions sustain engagement: a MOOC case study”. In: *IEEE Transactions on Learning Technologies* 10.4 (2016), pp. 475–487.

[31] Pradorn Sureephong et al. “The Effect of Non-Monetary Rewards on Employee Performance in Massive Open Online Courses.” In: *International Journal of Emerging Technologies in Learning* 15.1 (2020).

[32] Jihane Sophia Tahiri, Samir Bennani, and Mohammed Khalidi Idrissi. “diffMOOC: Differentiated Learning Paths Through the Use of Differentiated Instruction within MOOC.” In: *International Journal of Emerging Technologies in Learning* 12.3 (2017).

[33] Marshall P Thomas, Selen Türkay, and Michael Parker. “Explanations and interactives improve subjective experiences in online courseware”. In: *International Review of Research in Open and Distributed Learning* 18.7 (2017).

[34] Yimin Wen et al. “Consideration of the local correlation of learning behaviors to predict dropouts from MOOCs”. In: *Tsinghua Science and Technology* 25.3 (2019), pp. 336–347.

[35] Zheng Xie. “Modelling the dropout patterns of MOOC learners”. In: *Tsinghua Science and Technology* 25.3 (2019), pp. 313–324.

[36] Shengjun Yin et al. “Power of Attention in MOOC Dropout Prediction”. In: *IEEE Access* 8 (2020), pp. 202993–203002.

[37] Yafeng Zheng et al. “MOOC Dropout Prediction Using FWTS-CNN Model Based on Fused Feature Weighting and Time Series”. In: *IEEE Access* 8 (2020), pp. 225324–225335.

Appendices

Frequencies

Frequencies of ID_Curso

	Target	
ID_Curso	**Dropout**	**Success**
1	14	670
2	3	367
3	146	934
4	145	6155
5	18	260
6	38	1283
7	192	5784
8	54	1139
9	26	221
10	65	65
11	8	112
12	76	1061
13	34	430
14	100	623
17	0	28
18	129	618
19	56	355
21	3	80
25	40	44

Frequencies of Tem_Quiz?

	Target	
Tem_Quiz?	**Dropout**	**Success**
Não	157	402
Sim	990	19827

Frequencies of Número_questões_quiz

	Target	
Número_questões_quiz	**Dropout**	**Success**
0	157	402
2	69	253
3	129	4897
4	787	14675
5	4	2
6	1	0

Frequencies of Quiz_necessário?

	Target	
Quiz_necessário?	**Dropout**	**Success**
Não	130	405
Sim	1017	19824

Frequencies of Percentagem_necessária_para_quiz

	Target	
Percentagem_necessária_para_quiz	**Dropout**	**Success**
50	69	253
65	5	5
66	129	4897
70	17	1023
74	16	1030
75	752	12606
80	29	10

Frequencies of Resultado_quiz_(%)

	Target	
Resultado_quiz_(%)	**Dropout**	**Success**
0	58	13
25	5	0
50	49	186
66	11	179
67	1	28
75	109	2047
99	68	3763
100	565	14013

Frequencies of Tem_vídeo?

	Target	
Tem_vídeo?	**Dropout**	**Success**
Sim	1147	20229

Frequencies of Vídeo_necessário_(%)

	Target	
Vídeo_necessário_(%)	**Dropout**	**Success**
0	460	1837
90	423	16556
100	264	1836

Frequencies of Tem_inquérito?

	Target	
Tem_inquérito?	**Dropout**	**Success**
Não	7	26
Sim	1140	20203

Frequencies of Respondeu_inquérito?

	Target	
Respondeu_inquérito?	**Dropout**	**Success**
Não	1065	11280
Sim	82	8949

Frequencies of Nacionalidade

Nacionalidade	Target	
	Dropout	Success
0	3	43
Afeganistão	0	4
Alemanha	8	25
Angola	10	113
Brasil	14	140
Bulgária	0	4
Bélgica	1	2
Cabo Verde	8	174
Canadá	0	2
China	9	85
Colômbia	0	1
Congo	1	3
Dinamarca	0	1
Egipto	0	2
Equador	1	3
Espanha	2	33
Estados Unidos da América	1	9
França	17	15
Grã-Bretanha (Reino Unido, UK)	0	2
Grécia	0	1
Guiné	0	7
Guiné-Bissau	19	95
Holanda	0	11
Hungria	0	17
Itália	9	21
Jordânia	1	1
Lituânia	1	1
Macau/China	0	2
Moldávia	1	4
Moçambique	20	185
Noruega	2	7
Polónia	4	1
Portugal	1007	19082
Roménia	0	24
Rússia	0	2
Suíça	3	6
São Tomé e Príncipe	0	24
Sérvia	1	1
Síria	0	7
Timor-Leste	2	24
Ucrânia	1	31
Venezuela	1	11
África do Sul	0	1
Áustria	0	2

Frequencies of Pais_nascimento

	Target	
Pais_nascimento	**Dropout**	**Success**
Portugal	967	18301
0	5	49
Espanha	2	45
Moçambique	23	246
Brasil	20	229
França	22	61
Angola	10	182
África do Sul	1	37
São Tomé e Príncipe	2	42
Guiné-Bissau	19	125
Rússia	1	10
Suíça	4	73
Cabo Verde	9	209
Grã-Bretanha (Reino Unido, UK)	1	35
Ucrânia	7	105
China	8	73
Roménia	0	26
Liechtenstein	0	5
Macau/China	0	22
Moldávia	7	48
Bélgica	2	21
Alemanha	9	37
Timor-Leste	2	25
República do Senegal	0	3
Bielorússia	0	25
Colômbia	0	7
Porto Rico	0	4
Bulgária	0	4
Austrália	0	4
Estados Unidos da América	1	26
Venezuela	1	37
Afeganistão	0	4
República Democrática do Congo	1	3
Noruega	2	7
Polónia	5	5
Andorra	0	4
Canadá	1	4
Marrocos	0	1
Equador	1	3
Itália	8	20
Uzbequistão	0	2
Guiné	0	7
Síria	0	7
Egipto	0	8
Bangladesh	0	2
Holanda	0	5
Cuba	0	2

Frequencies of Pais_nascimento

	Target	
Pais_nascimento	**Dropout**	**Success**
Sérvia	1	1
Áustria	0	2
Índia	1	2
Luxemburgo	1	1
Israel	1	1
Arábia Saudita	1	1
Lituânia	1	1
Somália	0	2
Hungria	0	17
Dinamarca	0	1

Frequencies of Distrito-nascimento

	Target	
Distrito-nascimento	**Dropout**	**Success**
Ilha da Madeira (Madeira)	30	506
LISBOA	4	28
PORTO	0	2
Lisboa	546	10305
Alicante	0	5
Santarém	58	1049
0	108	1279
Setúbal	101	1967
Ilha de São Miguel (Açores)	15	280
Évora	18	357
Coimbra	23	433
Braga	10	96
Maputo	4	61
Porto	17	222
Lubango	1	11
Huambo	0	10
Bragança	3	50
Joanesburgo	1	13
Leiria	46	983
Água Grande	1	8
Guarda	5	219
Viseu	10	245
Faro	27	448
Castelo Branco	7	246
Portalegre	15	187
S. Tomense	0	4
Ilha Terceira (Açores)	5	162
Beja	11	244
Ilha do Faial (Açores)	1	29
Luanda	1	28
Ilha de Porto Santo (Madeira)	0	10
Aveiro	8	133
Ilha de Santiago	1	15
Nossa Senhora da Graça	1	3
Paraná	0	8
Suiça	0	4
Cabo Verde	2	35
Praia	0	12
Moçambique	3	10
Cacheu	0	4
Bissau	3	23
Zakarpathia	0	4
Guiné-Bissau	0	7
Joannesburg	0	4
Cabo verde	0	4
Suíça	0	4
Vila Real	3	31

Frequencies of Distrito-nascimento

	Target	
Distrito-nascimento	**Dropout**	**Success**
China	0	3
SERGIPE	0	2
Goiânia	0	3
Maramures	0	4
Macau	0	6
Andalucía	0	4
Guiné Bissau	0	14
Ilha de São Vicente	0	4
Maputo	0	6
Rio de Janeiro	2	12
Díli	0	4
ZheJiang	0	4
Chernivtsi	2	4
Praia	0	4
Blagoevgrad	0	4
Santiago	1	24
Viana do Castelo	2	24
Tombua	0	2
Ilha de S.Vicente	0	4
Ilha de São Jorge (Açores)	0	4
Guiné	0	7
Ilha do Pico (Açores)	0	6
Cabo Verde	0	2
Bafatá	0	4
Wesminster	0	4
LosPalos	2	12
Luanda	1	4
Ilha da Brava	0	4
Bolama	0	4
Nampula	0	4
Ritondo	1	6
Ingombotas	1	5
BELÉM	2	14
Angola	0	3
Ilha da Graciosa (Açores)	0	4
Dili	0	4
Praça de Titina Sila	0	4
Lubango-Angola	0	5
Huila	0	2
África do Sul	0	4
Teresina	0	2
Canadá	0	2
Ivano-Frankivsk	1	2
Ucrânia	0	2
Orleans	0	1
Ilha de Santa Maria (Açores)	1	5
Sª. da Graça - Praia	0	2
Paris	0	5

Frequencies of Distrito-nascimento

	Target	
Distrito-nascimento	**Dropout**	**Success**
Marrocos	0	1
Sambizanga	1	6
RIO GRANDE DO SUL	0	5
São Paulo	4	14
PRAIA	0	1
Beira	1	2
Taschkent	0	2
Guiné- Bissau	1	4
San Cristóbal	0	2
Namibe	0	4
Moçambique	0	4
Maianga	0	2
Dolj	0	2
Hamburgo	0	2
HUAMBO	1	2
Rio de Janeiro/RJ/Brasil	0	4
Minas Gerais	1	1
Espírito Santo	0	2
Mato Grosso do Sul	0	4
Bissau	0	2
São Tomé e Príncipe	0	2
Lobito	1	3
Caio	0	4
Huila - Lubango- Angola	0	1
Cabinda	1	1
Bissau, Guiné-Bissau	0	1
Hubei	0	1
Distrito urbano da Maianga	0	2
Locarno	0	2
S. Luis	0	2
Espírito Santo	0	2
Água grande	0	2
Sao tomé	0	2
Bahia	0	2
South Banat	1	1
ruijin，jiangxi province	0	1
Oppland	1	1
Gironde	0	2
Lombardia	0	2
Hochtaunuskreis	0	2
ANGOLA	0	5
Hubei Province	0	2
Guangdong province	0	2
Liaoning Province	0	2
Recife	0	2
California	1	3
Bayern	0	2
Bogota	0	3

Frequencies of Distrito-nascimento

	Target	
Distrito-nascimento	**Dropout**	**Success**
Bruxelas	0	2
TISWADI	0	2
Ilha das Flores (Açores)	0	1
Luxembourg	1	1
Lourenço Marques	0	1
Konstanz	1	2
Belem	0	2
Tel Aviv	1	1
Île-de-France	1	1
Vaud	1	1
Aseer	1	1
Shenzhen	1	1
Moldávia	1	1
Cologne	1	1
Pavia	1	1
Piacenza	1	1
Chisinau	1	1
Gansu	1	1
Bom Jardim da Serra	1	2
British Columbia	1	1
Bélgica	0	1
Satu-Mare	0	2
Aleppo	0	2
Dakar	0	1
Genebra	0	2
Benguela	0	1
Carabobo	0	4
Guiné Bissau	1	2
Mogadiscio	0	2
Kuanza Sul	0	2
França	0	2
São Paulo	0	2
Basel	0	2
Omsk	0	2
minas gerais	0	2
Sector Autonomo de Bissau	1	0
Bairro-Belém	2	0
Zavala	1	0
Gabú	1	0
Aquidauana	1	0
Machala	1	0
Begene	1	1
Mindelo	0	1
Bretagne	1	0
Thane	1	0
Guangdong	1	1
Shandong Province	0	1
Diemen	0	1

Frequencies of Distrito-nascimento

	Target	
Distrito-nascimento	**Dropout**	**Success**
Esslingen	1	0

Frequencies of País-residência

	Target	
País-residência	**Dropout**	**Success**
0	0	1
Alemanha	6	16
Angola	2	17
Brasil	0	18
Bélgica	1	2
Cabo Verde	1	12
China	1	5
Colômbia	0	1
Espanha	3	16
Estados Unidos da América	1	3
França	13	11
Grã-Bretanha (Reino Unido, UK)	0	2
Guiné-Bissau	3	12
Holanda	0	3
Hungria	0	17
Itália	4	13
Lituânia	1	1
Moçambique	6	24
Namíbia	0	1
Polónia	5	4
Portugal	1097	20046
Suíça	3	4

Frequencies of Sector_profissional_ (aluno)

Sector_profissional_ (aluno)	Target	
	Dropout	Success
0	21	255
Agricultores e trabalhadores qualificados da agricultura e pescas	1	5
Desconhecido/Não tem	781	15094
Especialistas das profissões intelectuais e científicas	24	296
Membros das Forças Armadas	8	73
Operadores de instalações e máquinas e trabalhadores da montagem	4	18
Operários, artífices e trabalhadores similares	2	33
Outra situação	182	3173
Pessoal administrativo e similares	25	291
Pessoal dos serviços e vendedores	38	396
Quadros superiores da Administração Pública, dirigentes e quadros superiores de empresa	26	199
Trabalhadores não qualificados	3	62
Técnicos e profissionais de nível intermédio	32	334

Frequencies of Condição_profissional_(mãe)

Condição_profissional_(mãe)	Target	
	Dropout	Success
0	21	255
Desconhecido/Não tem	66	923
Desempregado/a	69	1254
Doméstica/o	82	1082
Estudante	2	30
Outra situação	34	527
Reformado/a	53	713
Serviço militar	0	7
Trabalha para pessoas da família sem receber remuneração	3	26
Trabalha por conta de outrem	655	12894
Trabalha por conta própria - (como empregador)	83	1410
Trabalha por conta própria - independente (sem empregados)	79	1108

Frequencies of Sector_profissional_(mãe)

Sector_profissional_(mãe)	Target	
	Dropout	Success
0	21	255
Agricultores e trabalhadores qualificados da agricultura e pescas	8	124
Desconhecido/Não tem	181	2946
Especialistas das profissões intelectuais e científicas	90	2095
Membros das Forças Armadas	0	26
Operadores de instalações e máquinas e trabalhadores da montagem	5	88
Operários, artífices e trabalhadores similares	17	361
Outra situação	257	4309
Pessoal administrativo e similares	137	2815
Pessoal dos serviços e vendedores	116	1971
Quadros superiores da Administração Pública, dirigentes e quadros superiores de empresa	145	2421
Trabalhadores não qualificados	41	686
Técnicos e profissionais de nível intermédio	129	2132

Frequencies of Habilitações_literárias_(mãe)

Habilitações_literárias_(mãe)	Target	
	Dropout	Success
0	21	255
Desconhecido	49	381
Diploma de curso técnico superior profissional	4	45
Ensino Básico 1.º ciclo - 4.º ano de escolaridade (antiga 4ª classe)	43	860
Ensino Básico 2.º ciclo - 6.º ano de escolaridade (antigo 2ª ano liceal ou ciclo preparatório)	76	938
Ensino Básico 3.º ciclo - 9.º ano de escolaridade (antigo 5ª ano liceal ou ensino técnico)	147	2207
Ensino Médio	21	251
Ensino Pós-graduado - Doutoramento (pré-Bolonha)	14	341
Ensino Pós-graduado - Doutoramento 3º ciclo (Bolonha)	9	78
Ensino Pós-graduado - Mestrado (pré-Bolonha)	51	1268
Ensino Pós-graduado - Mestrado 2º ciclo (Bolonha)	10	138
Ensino Pós-secundário - Curso de especialização Tecnológica	12	173
Ensino Secundário - 12.º ano de escolaridade ou equivalente	254	5271
Ensino Superior - Bacharelato	62	1022
Ensino Superior - Licenciatura (Pré-Bolonha)	301	5798
Ensino Superior - Mestrado Integrado	13	221
Ensino superior - Licenciatura 1º ciclo (Bolonha)	39	713
Não sabe ler nem escrever	6	72
Sabe ler sem possuir o 4.º ano de escolaridade (antiga 4ª classe)	15	197

Frequencies of Condição_profissional_(pai)

Condição_profissional_(pai)	Target	
	Dropout	Success
0	21	255
Desconhecido/Não tem	83	1286
Desempregado/a	53	701
Doméstica/o	0	8
Estudante	1	11
Outra situação	52	850
Reformado/a	84	1175
Serviço militar	23	257
Trabalha para pessoas da família sem receber remuneração	2	5
Trabalha por conta de outrem	571	10954
Trabalha por conta própria - (como empregador)	170	3121
Trabalha por conta própria - independente (sem empregados)	87	1606

Frequencies of Sector_profissional_(pai)

Sector_profissional_(pai)	Target	
	Dropout	Success
0	21	255
Agricultores e trabalhadores qualificados da agricultura e pescas	19	320
Desconhecido/Não tem	142	2410
Especialistas das profissões intelectuais e científicas	62	1530
Membros das Forças Armadas	45	694
Operadores de instalações e máquinas e trabalhadores da montagem	43	680
Operários, artífices e trabalhadores similares	40	866
Outra situação	266	4449
Pessoal administrativo e similares	62	1284
Pessoal dos serviços e vendedores	129	2141
Quadros superiores da Administração Pública, dirigentes e quadros superiores de empresa	152	2584
Trabalhadores não qualificados	35	530
Técnicos e profissionais de nível intermédio	131	2486

Frequencies of Habilitações_literárias_(pai)

Habilitações_literárias_(pai)	Target	
	Dropout	Success
0	21	255
Desconhecido	72	733
Diploma de curso técnico superior profissional	8	73
Ensino Básico 1.º ciclo - 4.º ano de escolaridade (antiga 4ª classe)	70	1118
Ensino Básico 2.º ciclo - 6.º ano de escolaridade (antigo 2ª ano liceal ou ciclo preparatório)	64	1262
Ensino Básico 3.º ciclo - 9.º ano de escolaridade (antigo 5ª ano liceal ou ensino técnico)	154	2872
Ensino Médio	29	235
Ensino Pós-graduado - Doutoramento (pré-Bolonha)	19	466
Ensino Pós-graduado - Doutoramento 3º ciclo (Bolonha)	4	59
Ensino Pós-graduado - Mestrado (pré-Bolonha)	53	1304
Ensino Pós-graduado - Mestrado 2º ciclo (Bolonha)	5	151
Ensino Pós-secundário - Curso de especialização Tecnológica	22	286
Ensino Secundário - 12.º ano de escolaridade ou equivalente	272	5280
Ensino Superior - Bacharelato	64	948
Ensino Superior - Licenciatura (Pré-Bolonha)	218	4099
Ensino Superior - Mestrado Integrado	20	239
Ensino superior - Licenciatura 1º ciclo (Bolonha)	34	589
Não sabe ler nem escrever	3	64
Sabe ler sem possuir o 4.º ano de escolaridade (antiga 4ª classe)	15	196

```
|--- Respondeu_inquérito?<= 0.49
|   |--- Resultado_quiz_(%)<= 37.76
|   |   |--- Condição_profissional_(mãe)<= 5.65
|   |   |   |--- Condição_profissional_(mãe)<= 7.76
|   |   |   |   |--- Dropout
|   |   |   |--- Condição_profissional_(mãe)>  7.76
|   |   |   |   |--- Success
|   |   |--- Condição_profissional_(mãe)>  5.65
|   |   |   |--- Vídeo_necessário_(%) <= 0.05
…
|   |   |   |--- Vídeo_necessário_(%) >  0.05
|   |   |   |   |--- Dropout
|   |--- Resultado_quiz_(%)>  37.76
|   |   |--- Vídeo_necessário_(%) <= 21.87
|   |   |   |--- Resultado_quiz_(%)<= 77.38
|   |   |   |   |--- Condição_profissional_(mãe)<= 10.03
|   |   |   |   |   |--- Dropout
|   |   |   |   |--- Condição_profissional_(mãe)>  10.03
|   |   |   |   |   |--- Condição_profissional_(mãe)<= 13.10
|   |   |   |   |   |   |--- Dropout
|   |   |   |   |   |--- Condição_profissional_(mãe)>  13.10
|   |   |   |   |   |   |--- Success
|   |   |   |--- Resultado_quiz_(%)>  77.38
….
|   |   |--- Vídeo_necessário_(%) >  21.87
|   |   |   |--- Vídeo_necessário_(%) <= 99.98
…
|   |   |   |--- Vídeo_necessário_(%) >  99.98
…
|   |   |   |   |--- ID_Curso>  10.59
|   |   |   |   |   |--- Dropout
|--- Respondeu_inquérito?>  0.49
|   |--- ID_Curso<= 11.88
|   |   |--- Sector_profissional_ (aluno)<= 11.60
|   |   |   |--- Condição_profissional_(pai)<= 8.02
…
|   |   |   |   |--- Habilitações_literárias_(mãe)>  9.39
…
|   |   |   |--- Condição_profissional_(pai)>  8.02
…
|   |   |   |   |   |   |   |   |--- Success
|   |   |   |   |--- Condição_profissional_(mãe)>  10.72
…
|   |   |   |   |   |   |--- Success
|   |   |--- Sector_profissional_ (aluno)>  11.60
|   |   |   |--- Habilitações_literárias_(mãe)<= 9.77
|   |   |   |   |--- Habilitações_literárias_(mãe)<= 7.28
|   |   |   |   |   |--- Success
|   |   |   |   |--- Habilitações_literárias_(mãe)>  7.28
…
|   |   |   |   |   |   |--- Success
```

```
|   |   |   |--- Habilitações_literárias_(mãe)>  9.77
|   |   |   |   |--- Success
|   |--- ID_Curso>  11.88
|   |   |--- ID_Curso<= 16.18
|   |   |   |--- ID_Curso<= 13.27
|   |   |   |   |--- Success
|   |   |   |--- ID_Curso>  13.27
|   |   |   |   |--- Resultado_quiz_(%)<= 81.41
|   |   |   |   |   |--- Dropout
|   |   |   |   |--- Resultado_quiz_(%)>  81.41
...
|   |   |   |   |   |   |   |   |   |   |--- Dropout
|   |   |--- ID_Curso>  16.18
|   |   |   |--- ID_Curso<= 18.14
...
|   |   |   |   |   |   |--- Success
|   |   |   |--- ID_Curso>  18.14
|   |   |   |   |--- Success
```