Universidade de Évora - Escola de Ciências e Tecnologia

Mestrado em Engenharia Informática

Dissertação

Online Learning and Dropout

Rita Sofia Loureiro Guerreiro Cristina Leitão

Orientador(es) | Francisco Manuel Coelho
Irene Pimenta Rodrigues

Évora 2022

Universidade de Évora - Escola de Ciências e Tecnologia

Mestrado em Engenharia Informática

Dissertação

Online Learning and Dropout

Rita Sofia Loureiro Guerreiro Cristina Leitão

Orientador(es) | Francisco Manuel Coelho
Irene Pimenta Rodrigues

Évora 2022

A dissertação foi objeto de apreciação e discussão pública pelo seguinte júri nomeado pelo Diretor da Escola de Ciências e Tecnologia:

Presidente	Pedro Patinho (Universidade de Évora)
Vogais	Francisco Manuel Coelho (Universidade de Évora) (Orientador) Lígia Maria Ferreira (Universidade de Évora) (Arguente)

Évora 2022

To my family.

Acknowledgment

I would like to thank both teachers for all the help, meetings and guidance provided. The meetings and conversations were vital in inspiring me to think outside the box from multiple perspectives to form a comprehensive and objective critique. I would also like to extend my gratitude towards my family; their support was vital to help me accomplish and finish this chapter in my academic life.

Abstract

E-learning education is often more challenging than the traditional education method. In e-learning teaching, there is no face-to-face teaching, nor does it require instant feedback, the latter is provided during the learning process, making interactions more complex.

E-learning education is usually more challenging than traditional education because no presence or instant feedback is delivered during the learning process, making interactions more complex.

This dissertation aims to help identify students that can potentially be at risk of dropping or failing the e-learning course. To achieve that, we used Machine Learning (ML) classifiers, like Decision Tree (DT),Random Forest Classifier (RFC),Support Vector Classifier (SVC) and the AdaBoost Classifier (AC) to analyse the Open University Learning Analytics Dataset (OULAD). The duration of the courses was analysed in different fragments, so we could conclude if it would be possible to prevent a potential dropout beforehand.

We gathered that the best classifier in this study was AC, when analysing the **2/16** duration of the course with an accuracy of **64.6%**. The results also show that the SVC gives a very similar percentage for the accuracy when analysing the same period, **64.1%**. As a result of this study, we understand that it is preferable to analyse the data in the early stages of the course.

Ensino à distância e o seu abandono

A educação em e-learning geralmente é mais desafiadora do que o método de educação tradicional. No ensino de e-learning, o ensino presencial não, nem requer feedback instantâneo, este último é fornecido durante o processo de aprendizagem, tornando as interações mais complexas.

A educação em e-learning geralmente é mais desafiadora do que a educação tradicional porque nenhuma presença ou feedback instantâneo é fornecido durante o processo de aprendizagem, tornando as interações mais complexas.

Esta dissertação tem como objetivo auxiliar na identificação de alunos que podem estar em risco de abandono ou reprovar no curso de e-learning. Para isso, utilizamos classificadores Machine Learning (ML), como Decision Tree (DT), Random Forest Classifier (RFC), Support Vector Classifier (SVC) e AdaBoost Classifier (AC) para analisar o Open University Learning Analytics Dataset (OULAD). A duração dos cursos foi analisada em diferentes fragmentos, para que pudéssemos concluir se seria possível prevenir uma potencial evasão com antecedência.

Concluímos que o melhor classificador neste estudo foi o AC, ao analisar a duração **2/16** do curso com uma precisão de **64.6%**. Os resultados também mostram que o SVC fornece uma precisão muito semelhante ao analisar o mesmo período, **64.1%**. Como resultado deste estudo, entendemos que é preferível analisar os dados nas fases iniciais do curso.

Contents

List of Figures

List of Tables

List of Acronyms

URL	*Uniform Resource Locator*
ECT	Escola de Ciências e Tecnologia
UE	Universidade de Évora
k-NN	K-Nearest Neighbor
DT	Decision Tree
NB	Naive Bayes
NN	Neural Network
LMS	Learning Management System
DM	Data Mining
ML	Machine Learning
SL	Supervised Learning
OULAD	Open University Learning Analytics Dataset
RDS	Relational Database System
RFC	Random Forest Classifier
AC	AdaBoost Classifier
ACC	Accuracy
SVC	Support Vector Classifier
PM	Predictive Modeling
USL	Unsupervised Learning
RL	Reinforcement Learning
ANN	Artificial Neural Network

1

Introduction

With the appearance and the growth of the Internet, especially since the 1990s, there has been an increase in IT's use as a means of teaching and training. Since 1999, the word "e-learning" has been growing in educations systems, and with the growth of the Internet, e-learning systems have started to expand and started to gain a significant role in education.

But what is E-learning? E-learning is an online method of learning.

E-Learning systems have several benefits to individuals and companies. Individuals can improve their skills with courses and content, either unavailable in their location or whose site is too far away to allow travelling to attend in person. The e-learning system can help increase the population's education level. But dedication and engagement is one of the most important "tools" to succeed. One of the main factors to succeed in the E-learning system is the student engagement with the course and with other people.[SD10]

However, it is essential to mention that the e-Learning system can require considerable effort, especially when it comes to time management. As such, it is essential to create mechanisms that help teachers and students to prevent and identify, in early stages, possible cases of abandonment or failure.

Motivation

The main motivation behind this study is to help the students to be able to finish their courses. If the school have a mechanism to identify or predict if the student is at risk, the school can act accordingly to avoid that scenario.

What are the abandonment or failure rates for e-learning?

For an e-Learning student, time is precious, so the ability to manage the time correctly spent to investigate the available materials and research for multiple subjects is critical.

The student properties that are analyzed in the Open University Learning Analytics Dataset (OULAD) dataset are:

- Number of clicks in the materials available: This is the number of interactions the students have with the materials available during the course duration.
- Grades: Grades are how the students get scored at the end of the course. There are 4 different types of grades: Distinction, Pass, Fail and Withdrawn. For this study, the main focus is on the grades **Fail** and **Withdrawn** to find an answer to our case study.

There is also a personal motivation behind this topic since I do identify myself with the students that thought in dropping the course due to all the struggles that studying and working can bring.

Dissertation Structure

This dissertation starts with a brief introduction and shares the motivation for this study. The most common techniques used to solve similar problems are introduced in the chapter State of Art. The techniques mentioned in this dissertation are Data Mining (DM) and Machine Learning (ML).

After this brief technical introduction, a detailed description of the data and the database structure is presented in the chapter Open University Dropout Analysis.

In the chapter Data Experiences, the data is analysed against different machine learning classifiers and the results will be analysed. The classifiers used for this study are the Decision Tree (DT), Random Forest Classifier (RFC), Support Vector Classifier (SVC) and AdaBoost Classifier (AC).

Finally, in the chapter Conclusion and Future Work, there is a reflection about the conclusions and findings of this study and future work.

Case Study Question

The question for this study is whether it will be possible to predict student dropout or failure before the end of the course.

2

State of the Art

The following chapter describes previous projects in this area and some of the techniques used in their development. In addition, a brief description of Data Mining (DM) and how Machine Learning (ML) and these techniques can help with the question under study is also made.

2.1 Background

Learning Management System (LMS) are e-Learning platforms that enable and facilitate communication between teacher, students and the contents of the required course they are attending. This type of software has been growing in the last decades.

One example of LMS is Moodle, an open-source LMS that allows teacher-student communication in an e-Learning environment.

2.2 Important Techniques

In the development of a prediction system, some techniques are used in Data-Mining such as Machine learning classifiers for analysing the data are most used.

2.2.1 Data Mining

Data Mining (DM) is a task used to find patterns in large pieces of data in order to attempt to predict any future events based on given criteria. This type of analysis assumes that we can predict the outcomes based on patterns found in the data related to "older" events. DM has been used several times by many researchers to attempt to predict the dropout rate with several DM techniques.[ES14]

2.2.2 Machine Learning

Machine Learning (ML) is one of the subjects related to Artificial Intelligence that will help to develop the necessary prediction models which is the basis of our prediction and warning system. ML technique works by training an algorithm on a large data set, also called training set, in this study the data set is the Open University Learning Analytics Dataset (OULAD). In [DAM19], is mentioned that ML approach can be classified as Supervised Learning and Unsupervised Learning.

- Supervised Learning (SL): The use of labelled datasets defines this approach. Classification and Regression are the 2 types of problems for this approach.
- Unsupervised Learning (USL): This approach is defined by using ML algorithms to analyse and cluster unlabeled datasets, discovering the hidden patterns in the data. Clustering, Neural Networks and Anomaly detection are common algorithms used in this approach.
- Reinforcement Learning (RL): This approach is defined by making the decisions sequentially. The main difference between RL and SL is that in the latter, the training data contains the answer, so the model is trained with the correct answer. The same does not happen with reinforcement learning. In RL, the agent decides what to do to perform the task.

In this study the type of machine learning approach used is SL and the classifiers used to train the algorithm are: Decision Tree (DT), Random Forest Classifier (RFC), Support Vector Classifier (SVC) and AdaBoost Classifier (AC). In this approach, the main goal is to learn a model from the training data that is already labelled, that allow us to make predictions about unseen or future data. [Mir17]

2.3 Published Articles Related with the Case Study

In several articles, the authors tried to attempt to answer the question, "How to predict the abandonment of an LMS student".

In [TS15], for example, the dropout rate for the Open University (UK) is as high as 78%.

In China, the dropout rate for traditional learning is about 5%, while the dropout rate for E-learning is as high as 15-40%. This article shows how important it is to the schools to predict the dropout. To get the predictive model, the authors implemented machine learning techniques. "Bean and Rovai's models proposed that academic performance was directly related to dropping out, which has been validated by a majority of the empirical research".[TS15]

It is also mentioned, that it was necessary to split the research into small steps, in this case, four steps. "Step 1, Extract attribution data related to student dropouts from the information systems of online educational institutions, construct the training data set and feed the data into the dropout prediction model."

"Step 2, make use of the data to train the prediction models that were constructed based on machine learning methods such as the Artificial Neural Network (ANN), Decision Tree (DT) and Bayesian Network (BN) to derive the samples of the prediction model."

"Step 3, extract another section of data from the information systems for constructing a test data set and feed it into the actual samples of the prediction model previously generated."
"Step 4, make use of the prediction model samples to perform predictions on the test data set and evaluate the prediction results generated".

With all the collected data, they start creating the ANN and the DT. The ANN, is composed by three different layers, Input, Hidden and Output layer. It is important to explain that each layer correspond to each variable of the input attributes. The DT, is a tree structure, the root and internal nodes represent the input values of a certain attribution, the branch represents the output of the input value after the test, and the leaf node represents a specific category. [TS15]

In this article, we can admit that they have used persuasive techniques to predict dropout rates.

In the "European Journal of Open, Distance and E-learning", it is explained which techniques were used to analyze the data. In this study four Data-Mining methods were used, K-Nearest Neighbor (k-NN),DT, Naive Bayes (NB) and Neural Network (NN).[ES14]

There are other articles where the authors mention the use of different techniques such as Clustering Methods to analyze the outcome of the interactions between student and teacher and their relation to the student's final grade.[SH17]

In [DAIM15], it is explained how clustering can be useful to support the predictive process. Using the data logged from an Learning Management System (LMS), such as courses attempted, modules read, practice exam scores, student-student/student-teacher interactions and similar, patterns can be found using machine learning and pattern recognition techniques which can be used to indicate how a student is progressing.

Some authors also used Deep Learning models in their studies. In [DAIM15], the authors studied and analysed the data in different weeks and they concluded that the best approach would be Deep Learning based in the data they used, as this approach had the best Accuracy (ACC).

3

Open University Dropout Analysis

As mentioned before, in order to conduct this study the Open University Learning Analytics Dataset (OULAD) data was used and all our results are based on that data.

First we describe some core concepts of databases, and then we proceed to the description of the data in the OULAD.

Primary Key is a representation of a value that is unique for every record in a table, and the primary key column can't have null values.

Foreign Key is a way to link two different tables together. It is a field in one table that refers to a primary key.

Unique Constraint guarantee that all values in a column or a group of columns is unique or distinct from one another.

Composite Key is created with two or more attributes that together are unique.

It is also important to explain briefly the four types of cardinality that may exist in a Relational Database System (RDS):

One to many (1:M) This is the most common kind of relationship. A one-to-many relationship exists if only one of the related columns is a primary key or has a unique constraint.

Many to many (M:M) This kind of relationship occurs when each record of the first table can be related to one or more records on the second table and vice-versa. This type of relationship can be seen as a two one-to-many relationship which is linked by a 'linking table' or 'associate table'. The linking table links two tables by having fields which are the primary key of the other two tables.

One to one (1:1) This kind of relationship occurs when a single record that belongs to the first table is related to only one record on the second table and vice-versa.

Many to One (M:1) This kind of relationship occurs when multiple records that belongs to the first table are related to one record on the second table.

The type of cardinality in this database is 1 to 1 or Many, as depicted in figure 1.

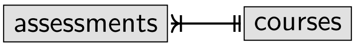

Figure 1: One course can have one or many assessments

3.1 Database Description

The OULAD database contains seven tables. Some represent entities while others represent associations and are grouped into three types:

Module Presentation Data related to the courses;

Student Activities Data related to the student activities and assessments;

Student Demographics Data related with the student details.

So now that we know the three kinds of information, we will describe in more detail the tables associate to each section.

The figure 2 shows the relationships between the tables and their attributes.

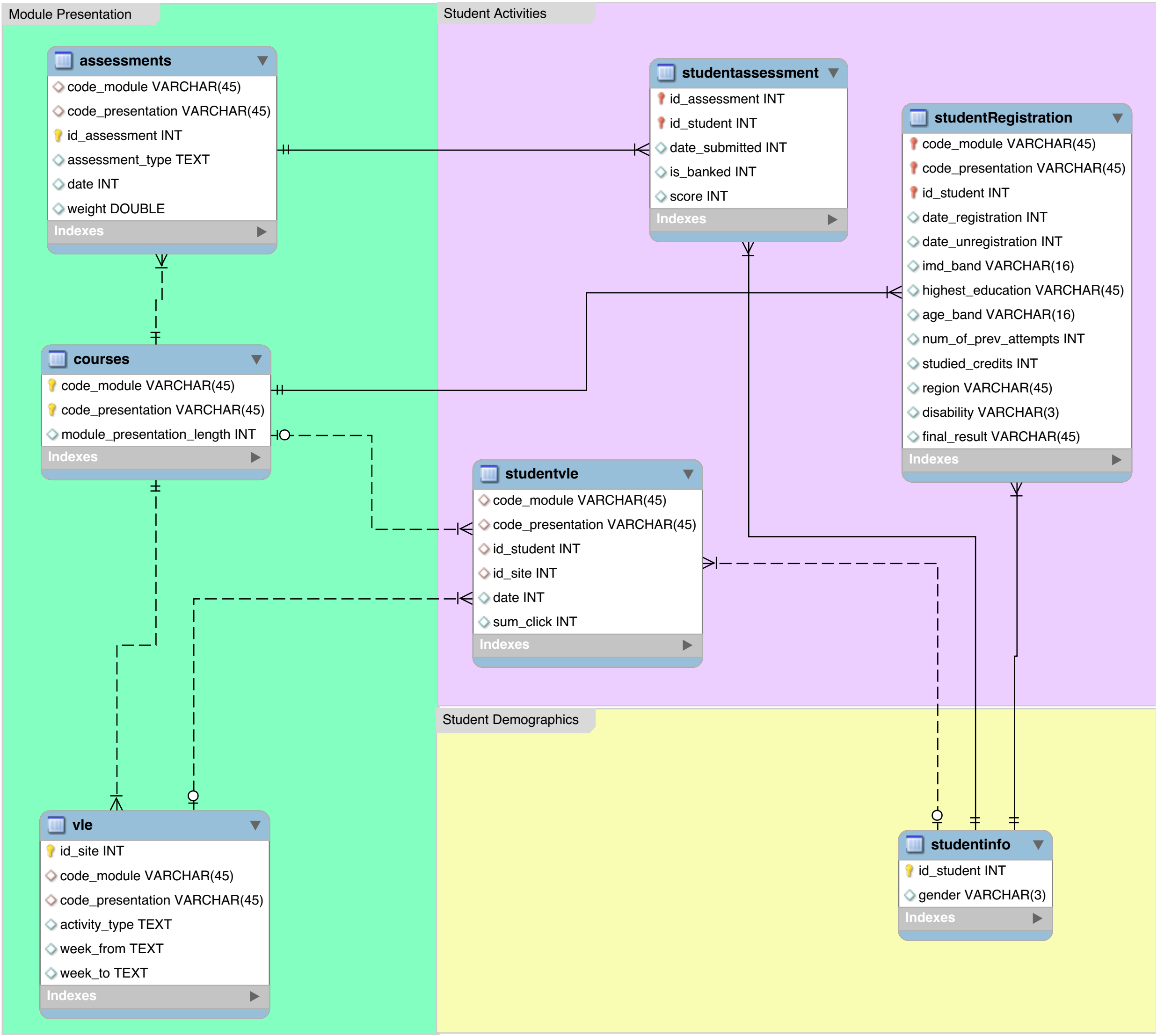

Figure 2: OULAD Enhanced Entity-Relationship Database

3.1.1 Module Presentation

The information about *Module Presentation* is represented by three tables detailed below.

The table Courses, contains a list of all the available modules and their presentations. This table has three attributes:

code_module is the identification code for the module, and it is one of the composite key in this table.

code_presentation is the time when the presentation starts.This consists of the year and “B” for the presentation starting in February and “J” for the presentation starting in October. It is also one of the composite key in this table

length it is a normal attribute and represents the length of the module-presentation in days.

It is important to explain that each combination of code_module and code_presentation will have one and only one length as shown in table 3.1.

code_module	code_presentation	length
AAA	2013J	268
AAA	2014J	269
BBB	2013J	268
BBB	2014J	262

Table 3.1: Courses: Each code_module can have different code_presentation values. These describe the year and semester.

The code_module and code_presentation are a composite key. These two attributes are foreign key in other tables, so they will be used to maintain the relationship between the tables. They are the connection between the table Courses to the tables Vle, Assessments, StudentVLE,StudentRegistration and StudentInfo.

The table Assessments contains all the information about the assessments in the *Module Presentation*. This table has six attributes described below.

code_module has been described previously. It is a foreign key in this table but a composite key on the table courses.

code_presentation has been described previously. This attribute is a foreign key in this table but is a composite key on the table courses

assessment is the identification code to which the assessment belongs, the primary key of this table.

assessment_type is the type of assessment. There are three types of assessments: Tutor Marked Assessment, Computer Marked Assessment and Final Exam.

date is the submission date of the assessments, and it is calculated based on starting date of the module-presentation, so on day one of the module-presentation the starting date is 0.

weight this attribute represents the weight of the assessment as a percentage of the final grade.

Table 3.2: Vle More than one id_site per combination of code_module, code_presentation and activity_type

id_site	code_module	code_presentation	activity_type
546943	AAA	2013J	resource
546712	AAA	2013J	oucontent
546678	AAA	2013J	oucontent
546998	AAA	2013J	resource
546961	AAA	2013J	resource
877256	AAA	2014J	resource
877423	AAA	2014J	dataplus
877088	AAA	2014J	dataplus

The table Assessments has 1:M relationship with StudentAssessment. So each assessment can be present in one or more fields on the table studentVle.

The last table to describe is the table Vle, this table stores the information about the available materials in the Vle. Information like the students interactions with the materials online. The table Vle contains the following attributes:

id_site it is the identification number of the material and is also a Primary Key;

code_module it is the identification for the code module. This attribute is a foreign key in this table but is a composite key on the table courses;

code_presentation it is the identification code of presentation. This attribute is a foreign key in this table but is a composite key on the table courses;

activity_type it is the role associated with the module material.

week_from refers to the week from which the material is planned to be used.

week_to refers to the week until which the material is planned to be used.

It is important to mention that we can have more than one id_site for the same combination of code_module and code_presentation for the same, activity_type as illustrated in the table 3.2.

3.1.2 Student Activities

The area of information related to the *Student Activities* contains three tables, and they are detailed below.

The table studentRegistration contains the information about the date when the student registered and unregistered in their courses. This table combines three foreign keys and other two attributes, see their description below.

code_module it is an identification code for a module also is foreign key int this table and part of the primary key on the table courses.

code_presentation the identification code of the presentation and as the previous attribute this is a foreign key in this table and a primary key on the table courses.

student is a unique identification number for the student, is an `foreign key` in this table and also a primary key on the table StudentInfo.

date_registration this attribute reflects the date of student's registration on the module presentation, this is the number of days measured relative to the start of the module-presentation (e.g. the negative value -30 means that the student registered to module presentation 30 days before it started).

date_unregistration is the date when student has done the un-registration from the module presentation, this is the number of days measured relative to the start of the module-presentation. Students, who completed the course, have this field empty. Students who unregistered have Withdrawal as the value of the `final_result`.

region identifies the geographic region, where the student lived while taking the course.

highest_education highest student education level on entry to the module presentation.

imd_band specifies the Index of Multiple Deprivation band of the place where the student lived during the course.

age_band refers to the band of the student's age.

num_of_prev_attempts is the number of times the student has attempted this module.

studied_credits the total number of credits for the modules the student is currently studying.

disability indicates whether the student has declared a `disability`.

final_result information about student's final result against the course.

It is important to mention that we can have one `student` associated with one or more combination of the `code_presentation` and `code_module`, as it is shown in the table 3.3.

The table `studentAssessment`, stores all the information related with the submission of the assessments and scores.

student this attribute is related with the table `StudentInfo` and is a `foreign key` in this table.

assessment this attribute refers to the identifier of the assessment and is a `foreign key` in this table.

date_submitted this attribute is the submission that of the assessment.

score this attribute refers to the student's score in a specific assessment.

is_banked this attribute is a stats flag to indicate if the student has been transferred from a previous presentation, ie, if the student attended this course previously.

Table 3.3: Student Registration: One Student with more than one Module Presentation

code_module	code_presentation	student	date_registration	date_unregistration
EEE	2013J	2698535	-74	-
CCC	2014B	2698535	-156	180

Table 3.4: Total of Students per final result

Final Results	Total	%
Distinction	3024	10.61%
Pass	12358	43.38%
Fail	6678	23.08%
Withdrawn	7168	22.92%

In the table 3.3, it is possible to see that there is more than one entry for each student according to the assessments that the student has done.

The table `StudentVle`,contains the information related with the student interaction with the material available.

`id_site` this attribute is a `foreign key` and also an identification number for the VLE material

`student` this attribute is a `foreign key` and it is a unique identification number for the student.

`code_module` this attribute is a `foreign key` and also an identification code for a module.

`code_presentation` this attribute is a `foreign key` and also the identification code of the module presentation.

`date` the date of student's interaction with the material measured as the number of days since the start of the module-presentation.

`sum_click` the number of times a student interacts with the material in that day.

3.1.3 Student Demographics

The last section to describe is the *Student Demographics* contains the table `StudentInfo`.

This table contains all the unique information related to the student.

`student` it is a unique identification number for the student.

`gender` refers to the gender information.

In the table `studentInfo`, `student`is the primary key and `code_module` and `code_presentation` are `foreign keys`. All the other column are attributes that can have null values.

3.2 Open University Data Summary

The table 3.4 shows a total of 29228 results, where 15382 are positive results (Pass or a Distinction), representing nearly 54.0% of the students. Which indicates that 46.0% failed or withdrawn.

In the figure 3.5 it is possible to observe that the students that Withdraw or Fail from the courses did not have as many interactions with the materials as the students that Passed or got Distinction in the courses DDD_2014J and AAA_2014J.

Table 3.5: Average of Clicks per Final Result for DDD_2014J and AAA_2014J courses

Final Result	DDD_2014J	AAA_2014J
Distinction	1758	3634
Pass	1359	1882
Fail	477	572
Withdrawn	293	925

4

Data Experiences

In this chapter, we will go through a sequence of experiments to understand the students' behaviours in the duration of the courses. The final goal is to evaluate a set of models of such relation using standard metrics such as score and accuracy.

Defining a numeric representation of student behaviour is necessary, particularly interaction with course resources. An example is the student's interactions with the available resources and how they behave in a certain period. It will be possible to predict if the students will drop from the course or not.

After analysing the initial database, it is clear that creating new tables that describe student activity and the final result will be necessary. Since there are many possible descriptions of "student activity" and "early weeks", some variations need to be investigated.

For this analysis, it was used the code in the appendix A, B and C.

For this study case, the positive classes are the "Withdraw" and "Fail", as the main goal of this study is to predict the student's dropout and abandonment.

The code in the appendix C will allow us to observe the confusion matrix and the score for the chosen classifier. The classifiers that will be part of this study are the following:

The Support Vector Classifier (SVC) main objective is to fit the data you provide, returning a “best fit” hyperplane that divides or categorizes your data. After getting the hyperplane, you can then feed some features to your classifier to see what the ”predicted” class is.

The Decision Tree (DT) is a supervised learning technique in Data Mining. This technique is used in classification problems. The main goal is to create a model able to predict the value of a target variable based on several input variables. There are two types of DT: Classification tree and Regression tree.

The Random Forest Classifier (RFC) an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean/average prediction (regression) of the individual trees. Random decision forests correct for decision tree’s habit of overfitting to their training set. Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees.

The AdaBoost Classifier (AC) adaptive because subsequent weak learners are tweaked in favour of those instances misclassified by previous classifiers. This is an algorithm where the weak learners will subsequently learn from their mistakes.

In order to obtain the Accuracy (ACC) value, it is necessary to calculate the accuracy for each model. The ACC is calculated based on the sum of all correct predictions (TP + TN) divided by the total number of the dataset (TP+TN+FN+FP), as you can see in the expression below:

$$ACC = \frac{TP + TN}{TP + TN + FN + FP}$$

The reason why it was decided to measure the ACC is because it is important to understand if the model has a good ACC or not to allow us to formulate an opinion and reach to a conclusion.

This study will be focused on analysing the same data in different time frames, so it is possible to find the best ACC.

The study will start by analysing the entire duration of the course divided into two equal parts, this will not be enough to do a proper analysis, so in order to have more suitable data for analysis, the duration of the course will need to be divided into smaller parts.

It is important to analyse the data in smaller parts and to observe the ACC of the classifier for this smaller parts, so it is possible for the teacher to predict if the student might be in danger of dropout the course or not.

Based on what has been previously described, to establish if the dataset enables to determinate the student approvals (Distinction, Pass), dropout (Withdrawn) or failed (Fail from the number of clicks (student behaviour) along the course four sets of experiences will be carried out:

1st set - 1/1, 2/2, 4/4, 8/8, 16/16 in this set of experiences: the final result is related to the total duration of the course, divided into 1, 2, 4, 8, and 16 equal length segments.

2nd set - 1/2, 2/4, 4/8, 8/16 in this set of experiences the data set is restricted to the first half of the total course duration.

3rd set - 1/4, 2/8, 4/16 in this set of experiences the focus will be on a quarter of the duration.

4th set - 1/8, 2/16 in this set of experiences the focus will be on the first 2 weeks of the duration.

Table 4.1: sum_click with the course materials for ten students. The 1of1 reference Clicks_1of1

course	student	1of1	final_result
AAA_014J	6516	2535	Pass
DDD_2013J	8462	565	Withdrawn
AAA_2013J	11391	836	Pass
CCC_2014J	23698	729	Pass
FFF_2013J	26247	367	Fail
BBB_2013B	27891	285	Withdrawn
FFF_2014B	42746	7948	Distinction
FFF_2013J	43958	2034	Withdrawn
DDD_2014J	46016	285	Withdrawn
BBB_2013J	63044	231	Fail

The main goal for these explorations is to find a model with the best ACC and understand if it is possible to predict in early stages if the student is following a pattern that can lead to a dropout or not.

This study will analyse the entire duration of the course first, but as predicting the dropout in a later stage will not bring any benefit to the student, there will also be analysis for the early moments of the course.

4.1 Experiences and Data Segments

4.1.1 1st Set - Experience 1 (1/1)

For this experience, it is necessary to analyse the student interactions with the materials during the entire course duration.

The dataset used for this experience has the following attributes:

course is also a Primary Key in this table represents the identification of the course (code_module and code_presentation together).

student is the Primary Key for this table and represent student's identification.

Clicks_1of1 this attribute will represent the sum of clicks that the student had during the duration of the course.

final_result refers to the final result of the student for the ModulePresentation.

Those attributes described above were chosen because they will be the most accurate attributes to study the data.

In the table 4.1, it is possible to observe the number of clicks that the student had with the materials during the entire duration of the course, but unfortunately that information is not sufficient to get an answer for our case study.

But it is possible to verify that the students that had an average of clicks below 500 have not concluded the course with success, as it is verified on the 4.1.

After preparing the data for the initial experience and having the dataset ready, it is time for the analysis. In the table 4.2 you can see the results obtained during this first experience.

Table 4.2: Results of the 1st Set - Experience 1 (1/1)

Models	Accuracy
Support Vector Classifier (SVC)	**77.4**
Decision Tree (DT)	73.9
Random Forest Classifier (RFC)	72.8
AdaBoost Classifier (AC)	77.3

After analysing the four classifiers, for this experience and using the data in the current state, the best Classifier would be the SVC.

The reason behind this is the fact that it has a slightly better score and this classifier had fewer less wrong predictions compared with the other classifiers.

Experience Results Summary:

Best Model : SVC

Best Model ACC : 77.4

But would this be the same if the data is analysed using smaller parts? That will the analysis to be performed in the next experience.

4.1.2 1st Set - Experience 2 (2/2)

The new dataset has the following attributes:

`course` is also a Primary Key in this table represents the identification of the course (code_module and code_presentation together).

`student` is the Primary Key for this table and represent students identification.

`Clicks_1of2` this attribute will represent the sum of clicks that the student had during the first half of the course

`Clicks_2of2` this attribute will represent the sum of clicks that the student had during the second half of the course.

`final_result` refers to the final result of the student for the ModulePresentation.

Those attributes described above were chosen because they will be the most accurate attributes to study the data.

With the selected attributes, it is possible to select two moments of the duration of the course and analyse the interaction of the student with the materials. The Table 4.3, shows the sum of clicks for ten students in the two distinct moments of the course.

The duration of the course was divided in half to see the students behaviours.

In the table 4.3, it is possible to see some students behaviour. The `student`"8462" had 0 interactions with the materials in the second half of the course, compared with the 565 interaction that he had in the first

Table 4.3: sum_click with the course materials for ten students of the course duration divided by 2. The 1of2 reference Clicks_1of2 and 2of2 reference the Clicks_2of2

course	student	1of2	2of2	final_result
AAA_2014J	6516	1347	1188	Pass
DDD_2013J	8462	565	0	Withdrawn
AAA_2013J	11391	612	224	Pass
CCC_2014J	23698	475	254	Pass
FFF_2013J	26247	367	0	Fail
BBB_2013B	27891	272	13	Withdrawn
FFF_2014B	42746	4210	3738	Distinction
FFF_2013J	43958	2034	0	Withdrawn
DDD_2014J	46016	284	1	Withdrawn
BBB_2013J	63044	231	0	Fail

Table 4.4: Analysis of the 1ST Set - Experience 2 (2/2)

Models	Accuracy
SVC	87.8
DT	82.3
RFC	86.7
AC	**87.9**

half of the course. The same happened for other students and some of them failed the course and others withdrawn.

The student"42746" had the opposite behaviour. This student had more than 4000 clicks in the first half and more than 3000 clicks in the second half and passed the course with Distinction.

So far, from what has been observed, the interactions of the students that withdrew from the courses dropped significantly in the second half of the course, from nearly 400 clicks to nearly 0 clicks.

After preparing the data for the initial experience and having the dataset ready, we started analysing the data. In the table 4.4 you can see the results obtained during this first exploration.

Experience Results Summary:

Best Model : AC

Best Model ACC : 87.9

The reason behind this is the fact that it has a slightly better score and this classifier had fewer less wrong predictions compared with the other classifiers.

But would this be the same if the data is analysed using smaller parts? That will the analysis to be performed in the next experience.

4.1.3 1st Set - Experience 3 (4/4)

For this experience, it is necessary to analyse the student interactions with the materials during the entire course duration, but in this experience the duration will be divided into 4 equal parts and the new dataset

Table 4.5: `sum_click` with the course materials for ten students of the course duration divided by 4. The 1of4 reference to `Clicks_1of4` and the same for the values up to 4of4.

course	student	1of4	2of4	3of4	4of4	final_result
AAA_2014J	6516	862	485	719	469	Pass
DDD_2013J	8462	446	119	0	0	Withdrawn
AAA_2013J	11391	447	165	38	186	Pass
CCC_2014J	23698	325	150	112	142	Pass
FFF_2013J	26247	367	0	0	0	Fail
BBB_2013B	27891	229	43	13	0	Withdrawn
FFF_2014B	42746	2529	1681	2497	1241	Distinction
FFF_2013J	43958	2016	18	0	0	Withdrawn
DDD_2014J	46016	248	36	1	0	Withdrawn
BBB_2013J	63044	161	70	0	0	Fail

Table 4.6: Analysis of the 1ST Set - Experience 3 (4/4)

Models	Accuracy
SVC	89.3
DT	86.0
RFC	91.1
AC	**91.3**

has the following attributes:

`course` is also a Primary Key in this table represents the identification of the course (code_module and code_presentation together).

`student` is the Primary Key for this table and represent student's identification.

`Clicks_1of4` this attribute will represent the sum of clicks that the student had during the first quarter of the course

`Clicks_2of4` this attribute will represent the sum of clicks that the student had during the second quarter of the course.

`Clicks_3of4` this attribute will represent the sum of clicks that the student had during the third quarter of the course.

`Clicks_4of4` this attribute will represent the sum of clicks that the student had during the fourth quarter of the course.

`final_result` refers to the final result of the student for the ModulePresentation.

The Table 4.5, shows the sum of clicks for ten students in the four distinct moments of the course. It is possible to observe in the table 4.5 that the students with a lower number of clicks in the `Clicks_1of4` and `Clicks_2of4` did not conclude the course with success, 2 students failed and 4 withdrawn from the course.

In the table 4.6 you can see the results obtained during this experience.

Experience Results Summary:

Best Model : AC

Best Model ACC : 91.3

But would this be the same if the data is analysed using smaller parts? That will the analysis to be performed in the next experience.

4.1.4 1st Set - Experience 4 (8/8)

For this experience, it is necessary to analyse the student interactions with the materials during the entire course duration, but in this experience the duration will be divided into 8 equal parts.

The dataset for this experience will be the same as the previous, but the focus will be on the following attribute:

`Clicks_1of8` this attribute will represent the sum of clicks that the student had during the 1/8 of the course.

`Clicks_2of8` this attribute will represent the sum of clicks that the student had during the 2/8 of the course.

`Clicks_3of8` this attribute will represent the sum of clicks that the student had during the 3/8 of the course.

`Clicks_4of8` this attribute will represent the sum of clicks that the student had during the 4/8 of the course.

`Clicks_5of8` this attribute will represent the sum of clicks that the student had during the 5/8 of the course.

`Clicks_6of8` this attribute will represent the sum of clicks that the student had during the 6/8 of the course.

`Clicks_7of8` this attribute will represent the sum of clicks that the student had during the 7/8 of the course.

`Clicks_8of8` this attribute will represent the sum of clicks that the student had during the 8/8 of the course.

The table 4.7, shows the sum of clicks for ten students in the eight distinct moments of the course. The duration of the course was divided by 8 to see the students behaviours, and it is possible to observe that the students with fewer clicks after the `Clicks_2of8` failed or withdrawn from the course.

In the table 4.8 you can see the results obtained during this exploration.

Experience Results Summary:

Best Model : RFC

Best Model ACC : 91.9

But would this be the same if the data is analysed using just the first half of the duration of the course? So the next experiences will be focus on the first 1/2 of the course and also splitting that into even smaller parts.

Table 4.7: `sum_click` with the course materials for ten students of the course duration divided by 8. The 1of8 reference `Clicks_1of8` and the same for the values up to 8of8.

`course`	`student`	1of8	2of8	3of8	4of8	5of8	6of8	7of8	8of8	`final_result`
AAA_2014J	6516	578	284	196	289	287	432	399	70	Pass
DDD_2013J	8462	338	108	97	22	0	0	0	0	Withdrawn
AAA_2013J	11391	329	118	47	118	3	35	93	93	Pass
CCC_2014J	23698	188	137	1	149	90	22	142	0	Pass
FFF_2013J	26247	367	0	0	0	0	0	0	0	Fail
BBB_2013B	27891	171	58	34	15	13	0	0	0	Withdrawn
FFF_2014B	42746	1421	1108	950	731	1355	1142	910	331	Distinction
FFF_2013J	43958	1121	895	7	11	0	0	0	0	Withdrawn
DDD_2014J	46016	174	74	18	18	0	1	0	0	Withdrawn
BBB_2013J	63044	93	68	48	22	0	0	0	0	Fail

Table 4.8: Analysis of the 1ST Set - Experience 4 (8/8)

Models	Accuracy
SVC	83.4
DT	81.6
RFC	**91.9**
AC	90.9

4.1.5 1st Set - Experience 5 (16/16)

For this experience, it is necessary to analyse the student interactions with the materials during the entire course duration, but in this experience the duration will be divided into 16 equal parts.

The dataset for this experience will be the same as the previous, but the focus will be on the following notation:

The notation `Clicks_nofm` represents the number of clicks in the n-th of m equal parts of the course duration. For example, `Clicks_5of8` is the column with the number of clicks in the 5 part of 8 equally divided parts of the course duration

The Table 4.9, shows the sum of clicks for ten students in the eight distinct moments of the course. The duration of the course was divided by 16 to see the students behaviours, and it is possible to observe that the students with fewer clicks after the `Clicks_2of8` failed or withdrawn from the course.

In the table 4.10 you can see the results obtained during this exploration.

Experience Results Summary:

Best Model : RFC

Best Model ACC : 91.4

But would this be the same if the data is analysed using just the first half of the duration of the course? So the next set of experiences will be focus on the first 1/2 of the course and also splitting that into even smaller parts.

Table 4.9: `sum_click` with the course materials for ten students of the course duration divided by 16. The 1of16 reference `Clicks_1of16` and the same for the values up to 16of16.

`course`	`student`	1of16	2of16	3of16	4of16	...	14of16	15of16	16of16	`final_result`
AAA_2014J	6516	296	282	124	160	...	233	68	2	Pass
DDD_2013J	8462	236	102	81	27	...	0	0	0	Withdrawn
AAA_2013J	11391	203	126	100	18	...	50	92	1	Pass
CCC_2014J	23698	166	22	14	123	...	3	0	0	Pass
FFF_2013J	26247	128	239	0	0	...	0	0	0	Fail
BBB_2013B	27891	149	29	35	49	...	0	0	0	Withdrawn
FFF_2014B	42746	647	774	724	384	...	609	230	101	Distinction
FFF_2013J	43958	647	474	453	442	...	0	0	0	Withdrawn
DDD_2014J	46016	109	65	41	33	...	0	0	0	Withdrawn
BBB_2013J	63044	40	53	68	0	...	0	0	0	Fail

Table 4.10: Analysis of the 1ST Set - Experience 5 (16/16)

Models	Accuracy
SVC	89.1
DT	86.0
RFC	**91.4**
AC	89.6

4.1.6 2nd Set - Experience 1 (1/2)

In the previous set of experience, an analysis was done against the entire duration of the courses, now in this set, the main focus is to analyse the first half of the duration of the course.

To achieve that, the dataset with the following attributes has been generated:

`course` is also a Primary Key in this table represents the identification of the course (code_module and code_presentation together).

`student` is the Primary Key for this table and represent student's identification.

`Clicks_1of2` this attribute will represent the sum of clicks that the student had during the first half of the course.

`final_result` refers to the final result of the student for the ModulePresentation.

The Table 4.11, shows the sum of clicks for ten students in the 1of2 of the course duration. It is also possible to observe that the students that had fewer interactions with the materials during the 1of2 end up failing or withdrawing the course.

In the table 4.12 you can see the results obtained during this exploration.

Experience Results Summary:

Best Model : AC

Best Model ACC : 69.6

Table 4.11: sum_click with the course materials for ten students during the Clicks_1of2 of the duration. The 1of2 reference Clicks_1of2.

course	student	1of2	final_result
AAA_2014J	6516	1347	Pass
DDD_2013J	8462	565	Withdrawn
AAA_2013J	11391	612	Pass
CCC_2014J	23698	475	Pass
FFF_2013J	26247	367	Fail
BBB_2013B	27891	272	Withdrawn
FFF_2014B	42746	4210	Distinction
FFF_2013J	43958	2034	Withdrawn
DDD_2014J	46016	284	Withdrawn
BBB_2013J	63044	231	Fail

Table 4.12: Analysis of the 2nd Set - Experience 1 (1/2)

Models	Accuracy
SVC	69.5
DT	66.3
RFC	65.9
AC	**69.6**

But would this be the same if the analysis of the data is done in even shorter periods? That will be the next experience. How will the models behave if the analysis is done using the first half divided into two even parts.

4.1.7 2nd Set - Experience 2 (2/4)

In the previous experience, it was possible to see the accuracy based just on the first half of the course duration, now for this experience, the main focus is to study the first 2/4 of the course duration.

In the table 4.13, it is possible to see that the students with a lower number of clicks in the second quarter did not successfully pass the course.

Table 4.13: sum_click with the course materials for ten students for the first quarter and second quarter of the course duration. The 1of4 reference Clicks_1of4 and 2of4 to Clicks_2of4.

course	student	1of4	2of4	final_result
AAA_2014J	6516	862	485	Pass
DDD_2013J	8462	446	119	Withdrawn
AAA_2013J	11391	447	165	Pass
CCC_2014J	23698	325	150	Pass
FFF_2013J	26247	367	0	Fail
BBB_2013B	27891	229	43	Withdrawn
FFF_2014B	42746	2529	1681	Distinction
FFF_2013J	43958	2016	18	Withdrawn
DDD_2014J	46016	248	36	Withdrawn
BBB_2013J	63044	161	70	Fail

Table 4.14: Analysis of the 2nd Set - Experience 2 (2/4)

Models	Accuracy
SVC	79.1
DT	71.9
RFC	75.8
AC	**79.4**

In the table 4.14 you can see the results obtained during this exploration.

In this experience, it is possible to verify that the best classifier is the AC, with the best Accuracy (ACC) and the best classifier score.

Experience Results Summary:

Best Model : AC

Best Model ACC : 79.4

4.1.8 2nd Set - Experience 3 (4/8)

In the previous experience, it was possible to see the accuracy based just on `Clicks_1of4` and `Clicks_2of4` of the course duration, now for this experience, the main focus is to study the first four `Clicks_1of8` of the course duration.

The dataset for this experience will have the following attributes:

`Clicks_1of8` this attribute will represent the sum of clicks that the student had during the 1/8 of the course.

`Clicks_2of8` this attribute will represent the sum of clicks that the student had during the 2/8 of the course.

`Clicks_3of8` this attribute will represent the sum of clicks that the student had during the 3/8 of the course.

`Clicks_4of8` this attribute will represent the sum of clicks that the student had during the 4/8 of the course.

The Table 4.15, shows the sum of clicks for ten students in the first `Clicks_4of8` distinct moments of the course. The duration of the course was divided by 8, but for this analysis the focus will be in the first `Clicks_4of8` to see the students behaviours, and it is possible to observe that the students with fewer clicks after the `Clicks_2of8` failed or withdrawn from the course.

In the table 4.16 you can see the results obtained during this exploration.

In this experience, it is possible to verify that the best classifier is the AC, with the best ACC and the best classifier score.

Experience Results Summary:

Table 4.15: `sum_click` with the course materials for ten students of the course duration divided by 8. The 1of8 reference `Clicks_1of8` and the same for the values up to 4of8.

`course`	`student`	1of8	2of8	3of8	4of8	`final_result`
AAA_2014J	6516	578	284	196	289	Pass
DDD_2013J	8462	338	108	97	22	Withdrawn
AAA_2013J	11391	329	118	47	118	Pass
CCC_2014J	23698	188	137	1	149	Pass
FFF_2013J	26247	367	0	0	0	Fail
BBB_2013B	27891	171	58	34	15	Withdrawn
FFF_2014B	42746	1421	1108	950	731	Distinction
FFF_2013J	43958	1121	895	7	11	Withdrawn
DDD_2014J	46016	174	74	18	18	Withdrawn
BBB_2013J	63044	93	68	48	22	Fail

Table 4.16: Analysis of the 2nd Set - Experience 3 (4/8)

Models	Accuracy
SVC	79.5
DT	72.4
RFC	79.1
AC	**80.5**

Best Model : AC

Best Model ACC : 80.5

4.1.9 2nd Set - Experience 4 (8/16)

In the previous experience, it was possible to see the accuracy based just on `Clicks_1of4` and `Clicks_2of4` of the course duration, now for this experience, the main focus is to study the first four `Clicks_1of8` of the course duration.

The dataset for this experience will have the following attributes:

`Clicks_1of16` this attribute will represent the sum of clicks that the student had during the 1/16 of the course.

`Clicks_2of16` this attribute will represent the sum of clicks that the student had during the 2/16 of the course.

`Clicks_3of16` this attribute will represent the sum of clicks that the student had during the 3/16 of the course.

`Clicks_4of16` this attribute will represent the sum of clicks that the student had during the 4/16 of the course.

`Clicks_5of16` this attribute will represent the sum of clicks that the student had during the 5/16 of the course.

Table 4.17: `sum_click` with the course materials for ten students of the course duration divided by 16. The 1of16 reference `Clicks_1of16` and the same for the values up to 8of16.

course	student	1of16	2of16	3of16	4of16	5of16	6of16	7of16	8of16	final_result
AAA_2014J	6516	296	282	124	160	176	20	65	224	Pass
DDD_2013J	8462	236	102	81	27	27	70	10	12	Withdrawn
AAA_2013J	11391	203	126	100	18	40	7	36	82	Pass
CCC_2014J	23698	166	22	14	123	0	1	10	139	Pass
FFF_2013J	26247	128	239	0	0	0	0	0	0	Fail
BBB_2013B	27891	149	29	35	49	2	32	15	0	Withdrawn
FFF_2014B	42746	647	774	724	384	631	319	591	140	Distinction
FFF_2013J	43958	647	474	453	442	7	0	11	0	Withdrawn
DDD_2014J	46016	109	65	41	33	0	18	18	0	Withdrawn
BBB_2013J	63044	40	53	68	0	0	48	22	0	Fail

Table 4.18: Analysis of the 2nd Set - Experience 4 (8/16)

Models	Accuracy
SVC	80.6
DT	74.4
RFC	**81.9**
AC	81.3

`Clicks_6of16` this attribute will represent the sum of clicks that the student had during the 6/16 of the course.

`Clicks_7of16` this attribute will represent the sum of clicks that the student had during the 7/16 of the course.

`Clicks_8of16` this attribute will represent the sum of clicks that the student had during the 8/16 of the course.

The Table 4.17, shows the sum of clicks for ten students in the first `Clicks_8of16` distinct moments of the course. The duration of the course was divided by 16, but for this analysis the focus will be in the first `Clicks_8of16` to see the students behaviours, and it is possible to observe that the students with fewer clicks after the `Clicks_4of16` failed or withdrawn from the course.

In the table 4.18 you can see the results obtained during this exploration.

In this experience, it is possible to verify that the best classifier is the RFC, with the best ACC and the best classifier score.

Experience Results Summary:

Best Model : RFC

Best Model ACC : 81.9

Table 4.19: Analysis of the 3rd Set - Experience 1 (1/4)

Models	Accuracy
SVC	64.3
DT	62.2
RFC	61.8
AC	**64.4**

4.1.10 3rd Set - Experience 1 (1/4)

In the previous set of experience, an analysis was done against half of the duration of the courses, now in this set, the main focus is to analyse a quarter of the duration of the course.

In the table 4.13, it is possible to see that the students with a lower number of clicks in the second quarter did not successfully pass the course.

For this analysis, the same that have been in previous experiences, but the main focused was the use of the `Clicks_1of4`.

In the table 4.19 you can see the results obtained during this exploration.

In this experience, it is possible to verify that the best classifier is the AC, with the best ACC and the best classifier score.

Experience Results Summary:

Best Model : AC

Best Model ACC : 64.4

4.1.11 3rd Set - Experience 2 (2/8)

In the previous experience, was possible to see the accuracy based on `Clicks_1of4` of the course duration, now for this exploration, the main focus is to study the first two `Clicks_1of8` of the course duration.

The dataset for this experience has the following attributes:

`Clicks_1of8` this attribute will represent the sum of clicks that the student had during the 1/8 of the course.

`Clicks_2of8` this attribute will represent the sum of clicks that the student had during the 2/8 of the course.

The Table 4.20, shows the sum of clicks for ten students in the first `Clicks_2of8` distinct moments of the course. The duration of the course was divided by 8, but for this analysis the focus will be in the first `Clicks_2of8` to see the students behaviours, and it is possible to observe that the students with fewer clicks after the `Clicks_1of8` failed or withdrawn from the course.

In the table 4.21 you can see the results obtained during this exploration.

In this experience, it is possible to verify that the best classifier is the SVC, with the best ACC and the best classifier score.

Table 4.20: `sum_click` with the course materials for ten students of the course duration divided by 8. The 1of8 reference `Clicks_1of8` and the same for the values up to 2of8.

`course`	`student`	1of8	2of8	`final_result`
AAA_2014J	6516	578	284	Pass
DDD_2013J	8462	338	108	Withdrawn
AAA_2013J	11391	329	118	Pass
CCC_2014J	23698	188	137	Pass
FFF_2013J	26247	367	0	Fail
BBB_2013B	27891	171	58	Withdrawn
FFF_2014B	42746	1421	1108	Distinction
FFF_2013J	43958	1121	895	Withdrawn
DDD_2014J	46016	174	74	Withdrawn
BBB_2013J	63044	93	68	Fail

Table 4.21: Analysis of the 3rd Set - Experience 2 (2/8)

Models	Accuracy
SVC	**69.5**
DT	62.9
RFC	64.9
AC	69.4

Experience Results Summary:

Best Model : SVC

Best Model ACC : 69.5

4.1.12 3rd Set - Experience 3 (4/16)

In the previous experience, was possible to see the accuracy based on `Clicks_1of8` and `Clicks_2of8` of the course duration, now for this experience, the main focus is to study the first 4 `Clicks_1of16` of the course duration.

The dataset for this experience has the following attributes:

`Clicks_1of16` this attribute will represent the sum of clicks that the student had during the 1/2 of the course.

`Clicks_2of16` this attribute will represent the sum of clicks that the student had during the 2/16 of the course.

`Clicks_3of16` this attribute will represent the sum of clicks that the student had during the 3/16 of the course.

`Clicks_4of16` this attribute will represent the sum of clicks that the student had during the 4/16 of the course.

The Table 4.22, shows the sum of clicks for ten students in the first `Clicks_4of16` distinct moments of the course. The duration of the course was divided by 16, but for this analysis the focus will be in the

Table 4.22: sum_click with the course materials for ten students of the course duration divided by 16. The 1of16 reference Clicks_1of16 and the same for the values up to 4of16.

course	student	1of16	2of16	3of16	4of16	final_result
AAA_2014J	6516	296	282	124	160	Pass
DDD_2013J	8462	236	102	81	27	Withdrawn
AAA_2013J	11391	203	126	100	18	Pass
CCC_2014J	23698	166	22	14	123	Pass
FFF_2013J	26247	128	239	0	0	Fail
BBB_2013B	27891	149	29	35	49	Withdrawn
FFF_2014B	42746	647	774	724	384	Distinction
FFF_2013J	43958	647	474	453	442	Withdrawn
DDD_2014J	46016	109	65	41	33	Withdrawn
BBB_2013J	63044	40	53	68	0	Fail

Table 4.23: Analysis of the 3rd Set - Experience 3 (4/16)

Models	Accuracy
SVC	70.3
DT	63.3
RFC	69.6
AC	**70.9**

first Clicks_4of16 to see the students behaviours and it is possible to observe that the students with less clicks after the Clicks_2of16 failed or withdrawn from the course.

In the table 4.23 you can see the results obtained during this exploration.

In this experience, the best classifier is the AC, with the best ACC and the best classifier score.

Experience Results Summary:

Best Model : AC

Best Model ACC : 70.9

4.1.13 4th Set - Experience 1 (1/8)

In the previous set of experiences, it was possible to see the accuracy based on first quarter of the course duration, now for this experience, the main focus is to study the Clicks_1of8 of the course duration.

The dataset for this experience has the following attributes:

Clicks_1of8 this attribute will represent the sum of clicks that the student had during the 1/8 of the course.

The Table 4.24, shows the sum of clicks for ten students in the first Clicks_1of8 distinct moments of the course.

The duration of the course was divided by 8, but for this analysis the focus will be in the Clicks_1of8

Table 4.24: `sum_click` with the course materials for ten students of the course duration in the first 1of8. The 1of8 reference `Clicks_1of8`.

course	student	1of8	final_result
AAA_2014J	6516	578	Pass
DDD_2013J	8462	338	Withdrawn
AAA_2013J	11391	329	Pass
CCC_2014J	23698	188	Pass
FFF_2013J	26247	367	Fail
BBB_2013B	27891	171	Withdrawn
FFF_2014B	42746	1421	Distinction
FFF_2013J	43958	1121	Withdrawn
DDD_2014J	46016	174	Withdraw

Table 4.25: Analysis of the 4th Set - Experience 1 (1/8)

Models	Accuracy
SVC	**62.7**
DT	60.4
RFC	60.0
AC	62.5

to see the students behaviours, and it is possible to observe that the students with fewer clicks during this period failed or withdrawn from the course.

In the table 4.25 you can see the results obtained during this exploration.

In this experience, the best classifier is the AC, with the best ACC and the best classifier score.

Experience Results Summary:

Best Model : AC

Best Model ACC : 62.7

4.1.14 4th Set - Experience 2 (2/16)

In the previous experiences, was possible to see the accuracy based on `Clicks_1of8` of the course duration, now for this experience, the main focus is to study the `Clicks_2of16` of the course duration.

The dataset for this experience has the following attributes:

`Clicks_1of16` this attribute will represent the sum of clicks that the student had during the 1/16 of the course.

`Clicks_2of16` this attribute will represent the sum of clicks that the student had during the 2/16 of the course.

The Table 4.26, shows the sum of clicks for ten students in the `Clicks_2of16` distinct moments of the course. The duration of the course was divided by 16, but for this analysis the focus will be in the first

Table 4.26: sum_click with the course materials for ten students of the course duration in the first 2of16. The 1of16 reference Clicks_1of16 and 2of16 reference Clicks_2of16

course	student	1of16	2of16	final_result
AAA_2014J	6516	296	282	Pass
DDD_2013J	8462	236	102	Withdrawn
AAA_2013J	11391	203	126	Pass
CCC_2014J	23698	166	22	Pass
FFF_2013J	26247	128	239	Fail
BBB_2013B	27891	149	29	Withdrawn
FFF_2014B	42746	647	774	Distinction
FFF_2013J	43958	647	474	Withdrawn
DDD_2014J	46016	109	65	Withdrawn
BBB_2013J	63044	40	53	Fail

Table 4.27: Analysis of the 4th Set - Experience 1 (2/16)

Models	Accuracy
SVC	64.1
DT	57.8
RFC	59.9
AC	**64.6**

Clicks_1of16 to see the students behaviours, and it is possible to observe that the students with fewer clicks after the Clicks_2of16 failed or withdrawn from the course.

In the table 4.27 you can see the results obtained during this exploration.

In this experience, the best classifier is the AC, with the best ACC and the best classifier score.

Experience Results Summary:

Best Model : AC

Best Model ACC : 64.6

4.2 Summary and Final Observations

As it has been mentioned earlier in this dissertation, the main focus of this study is to try to understand if it is possible or not to predict the student dropout from the courses with the data available.

In table 4.28 it is possible to observe that the experience 8/8 had the best results in terms of Accuracy (ACC) and the model Random Forest Classifier (RFC) was the best classifier for this data, this is due to the fact that the data was divided in 8 smaller parts.

In table 4.29 we can observe the different accuracy between the classifiers using just the first half of the duration of the course. It is also possible to observe that the 8/16 experience had the best results in terms of ACC and the model RFC was the best classifier for this data.

In table 4.30 we can observe the different accuracy between the classifiers using 1/4 of the duration of the course. It is also possible to observe that the 4/16 experience had the best results in terms of ACC and

Table 4.28: Accuracy for the 1st Set of Experiences. It is possible to observe the different accuracy between the classifiers in the previous experiences that studied the entire duration of the course.

	SVC	Decision Tree	Random Forest	AdaBoost
1/1	77.4	73.9	72.8	77.3
2/2	87.8	82.3	86.7	87.9
4/4	89.3	86.0	91.1	91.3
8/8	83.4	81.6	**91.9**	90.9
16/16	89.1	86.0	91.4	89.6

Table 4.29: ACC for the 2nd Set of Experiences.

	SVC	Decision Tree	Random Forest	AdaBoost
-/1	-	-	-	-
1/2	69.5	66.3	65.9	69.6
2/4	79.1	71.9	75.8	79.4
4/8	79.5	72.4	79.1	80.5
8/16	80.6	74.4	**81.9**	81.3

the model AdaBoost Classifier (AC) was the best classifier for this data.

Table 4.30: ACC for the 3rd Set of Experiences.

	SVC	Decision Tree	Random Forest	AdaBoost
-/1	-	-	-	-
-/2	-	-	-	-
1/4	64.3	62.2	61.8	64.4
2/8	69.5	62.9	64.9	69.4
4/16	70.3	63.3	69.6	**70.9**

The table 4.31 shows the first 1/8 of the duration of the course. It is also possible to observe that the 2/16 experience had the best results in terms of ACC and the model AC was the best classifier for this data.

The table 4.32 illustrates the entire duration of the course divided into 4 parts. It is also possible to observe that the 4/4 experience had the best results in terms of ACC and the model AC was the the best classifier for this data.

The table 4.33 illustrates the entire duration of the course divided into 8 parts. It is also possible to observe that the 8/8 experience had the best results in terms of ACC and the model RFC was the best classifier for this data.

The 4.34 illustrates the entire duration of the course divided into 16 parts. It is also possible to observe that the 16/16 experience had the best results in terms of ACC and the model RFC was the best classifier for this data.

Of course, we can't base our findings using the values present on the table: 4.28 as the best results because they reflect the entire duration of the course and that is not the purpose of this dissertation. The purpose of this dissertation is trying to predict the dropout in early stages of the course, so the teacher and the student can avoid the dropout.

In the previous tables, it is possible to understand that the best accuracy is **91.9** for the classifier AC, when analysing the data split into 8 identical parts (8/8). However, that will not help to achieve the main goal

Table 4.31: ACC for the 4th Set of Experiences.

	SVC	Decision Tree	Random Forest	AdaBoost
-/1	-	-	-	-
-/2	-	-	-	-
-/4	-	-	-	-
1/8	62.7	60.4	60.0	62.5
2/16	64.1	57.8	59.9	**64.6**

Table 4.32: ACC for the duration divided by 4

	SVC	Decision Tree	Random Forest	AdaBoost
1/4	64.3	62.2	61.8	64.4
2/4	79.1	71.9	75.8	79.4
4/4	89.3	86.0	91.1	**91.3**

of this study.

In the 4.31, which reflects the values for the 4th set of this experience(which is focus on the smaller parts of the course duration), it is possible to observe that the preferable way to predict the dropout is by using the 2/16 of the duration of the course and to use the **AdaBoost Classifier** which obtained an accuracy of **64.6%**.

Table 4.33: ACC for the duration divided by 8

	SVC	Decision Tree	Random Forest	AdaBoost
1/8	62.7	60.4	60.0	62.5
2/8	69.5	62.9	64.9	69.4
4/8	79.5	72.4	79.1	80.5
8/8	89.4	86.1	**91.9**	90.9

Table 4.34: ACC for the duration divided by 16

	ACC Support Vector Classifier (SVC)	ACC Decision Tree (DT)	ACC RFC	ACC AC
2/16	64.1	57.8	59.9	64.6
4/16	70.3	63.3	69.6	70.3
8/16	80.6	74.4	81.9	81.3
16/16	89.1	86.0	**91.4**	89.6

5

Conclusion and Future Work

In the previous chapter, after analysing the data, it was possible to conclude the best result for our study was the value of **64.6%** accuracy for the **2/16** of the duration of the course. The best fit classifier for this duration was the AdaBoost Classifier (AC). The above result is not the best Accuracy (ACC) against the different models and time frames but it is the best fit for our study, which is to be able to predict the dropout early as possible.

There are some factors that will play an important role in this first 2 weeks as well, ie, the students might be more interesting in the beginning of the course than in a later stage but we also understand these are very important weeks for the progress of the student. To achieve that, teacher(s)/school(s) should analyse the data using the AC, as early and possible and observe the interactions of the students with the materials available and to understand which student can be potentially in risk of dropout. We can assume this would be an excellent asset for the teachers and schools.

The suggestion for teacher(s)/school(s) is to start analysing the interactions the students have with the materials as soon as possible, as it is possible to start building a predictive model. In this study, we analysed all the courses together, but it would be interesting to do different analyses for future work. For future work or research could be in divided in two different ways: (a) it would be interesting to see the results per course; (b) focus the study on a group of 10 students and see their behaviour against the different courses they enrolled. These two different approaches would help to understand if the results would be the same or if they would differ on the time frame or best classifier.

A

MySQL - Function to Analyse the Data in Different Time Frames

The following code describes how the initial data was organized and split into the different timeframes.

Listing A.1: Function to analyse the data in different time frames.

```
DELIMITER $$
CREATE DEFINER=`root`@`localhost` FUNCTION `CreateSQLForAnalysis`(splits int)
    RETURNS TEXT CHARSET utf8
    DETERMINISTIC
BEGIN

SET @clickSelects = '';
SET @p1 = -1;

_loop: LOOP
```

```
   SET @p1 = @p1 + 1;
   IF @p1 < splits THEN
              SET @clickSelects = CONCAT(@clickSelects,'SUM(case when ',
              CONCAT('`studentactivityanalysis`.`activity_date` >= (`
                 studentactivityanalysis`.`module_presentation_length` / ', splits,
                  ') * ', @p1, ' '),
              'and ',
              CONCAT('`studentactivityanalysis`.`activity_date` <= (`
                 studentactivityanalysis`.`module_presentation_length` / ', splits,
                  ') * (', @p1,' + 1) '),
              'then `studentactivityanalysis`.`sum_click` ',
              'else 0 end',
              CONCAT(') AS `sum_', @p1+1, '`'),',');
   ITERATE _loop;
   END IF;
   LEAVE _loop;
END LOOP _loop;


set @query = CONCAT('SELECT ',
       '`studentactivityanalysis`.`id_student` AS `id_student`,',
   @clickSelects,
       '`studentactivityanalysis`.`code_module` AS `code_module`,',
       '`studentactivityanalysis`.`code_presentation` AS `code_presentation`,',
       '`studentactivityanalysis`.`final_result` AS `final_result` ',
       'FROM `studentactivityanalysis` ',
       'GROUP BY ',
       '`studentactivityanalysis`.`id_student`,',
       '`studentactivityanalysis`.`code_module` ,',
       '`studentactivityanalysis`.`code_presentation`,',
       '`studentactivityanalysis`.`final_result`;');

return @query;

END$$
DELIMITER ;
```

B

MySQL - Call Function

Listing B.1: Call the function that will return the duration of the course split in x times.

```
CREATE DEFINER=`root`@`localhost` PROCEDURE `GetAnalysisByTimeSplit`(
       IN timeSplit int
)
BEGIN
SET @sql = CreateSQLForAnalysis(timeSplit);

PREPARE stmt FROM @sql;
EXECUTE stmt;
DEALLOCATE PREPARE stmt;

END

CALL GetAnalysisByTimeSplit(numberOfTimestoSplit)
```

C

Python - Models Analysis

The following code describes how we were able to obtain the scores for the confusion matrix for the data in analysis.

The "USED_COLUMNS" columns represent the X value, which is the observed value. Also meaning the known value. After having the X defined, it is necessary to determine which parameter would be the Y, hidden value, also meaning the predicted value. For this first experience, the value Y is represented by the column "final_result". After defining the X and Y, it is necessary to specify which values would be the "Positive_Classes". For this study case, the best values would be "Withdraw" and "Fail", as the main goal of this study is to predict the student's dropout. The value Y was defined as a binary value, and that was set up by the parameter True or False. After defining the X and Y and setting the BIN_CLASSES as True or False, the next step was to compute the Linear Regression.

Listing C.1: Code that return the score for each model

```python


from sklearn.model_selection import train_test_split
```

```
USED_COLUMNS = # ['Clicks']

X = df[USED_COLUMNS].to_numpy().reshape(-1, len(USED_COLUMNS))
y_cat = df['final_result']
POSITIVE_CLASSES =['Withdrawn','Fail']
y_bin = np.array([1 if yi in POSITIVE_CLASSES else 0 for yi in y_cat])
BIN_CLASSES = True
y = y_bin if BIN_CLASSES else y_cat

X_train, X_test, y_train, y_test = train_test_split(X, y)

if BIN_CLASSES:
    #
    # Make and compute the score of a LinearRegression model
    #
    from sklearn.linear_model import LinearRegression

    linreg_model = LinearRegression(normalize=True)
    linreg_model.fit(X_train, y_train)
    score = linreg_model.score(X_test, y_test)
    print(score)


from sklearn.metrics import confusion_matrix, plot_confusion_matrix
from sklearn.metrics import multilabel_confusion_matrix
import matplotlib.pyplot as plt
from sklearn.metrics import accuracy_score

def study_model(m, X_train, X_test, y_train, y_test):
    """Train, evaluate and print a model"""
    #
    # FIT: Find the best values of the model parameters
    # using _train data
    #
    m.fit(X_train, y_train)
    #
    # SCORE: Evaluate the performance of the model
    # using _test data
    #
    score = m.score(X_test, y_test)

    print(f"\n{20*'='}\nModel: {m}\n\tscore: {score}")
    plot_confusion_matrix(m, X_test, y_test)
    plt.show()
    return m, score

from sklearn.svm import SVC
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.ensemble import AdaBoostClassifier
from sklearn.metrics import accuracy_score
#
# ==============================================================================
#
# Models to evaluate
#
# SETUP:
#
MODELS = [
    SVC(),
    DecisionTreeClassifier(),
    RandomForestClassifier(),
    AdaBoostClassifier(),
]
#
# ==============================================================================
#
# "Study" each model from above (fit, score and print).
# Also: find the best one.
#
best_score = -1
best_model = None
for model in MODELS:
    m, score = study_model(model, X_train, X_test, y_train, y_test)
    if score > best_score:
        best_model = m
        best_score = score
#
# ==============================================================================
#
print(f"\n\nThe best model is {best_model}\n\twith score {best_score}.")
```

Bibliography

[DAIM15] Ashish Dutt, Saeed Aghabozrgi, Maizatul Akmal Binti Ismail, and Hamidreza Mahroeian. Clustering algorithms applied in educational data mining. *International Journal of Information and Electronics Engineering*, 5:112–116, 2015. [Online; accessed 7-January-2018].

[DAM19] Dr. Vikram Bali Deepti Aggarwal and Sonu Mittal. An insight into machine learning techniques for predictive analysis and feature selection. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, 8:345, 2019. [Online; accessed 11-December -2021].

[ES14] Yukselturk Erman and Ozekes Serhat. Predicting dropout student: An application of data mining methods in an online education program. *European Journal of Open, Distance and e-Learning*, 17(1):123–123, 2014.

[Gua12] Alex Guazzelli. Predictive modeling techniques. *IBM Developer*, 2012. [Online; accessed 7-January-2018].

[Mir17] Sebastian Raschka & Vahid Mirjalili. *Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow.*, volume 2. Packet Publishing ltd, 2017.

[SD10] Suksan Suppasetseree and Nootprapa Dennis. The use of moodle for teaching and learning english at tertiary level in thailand. *THE INTERNATIONAL JOURNAL OF THE HUMANITIES*, 8:23, 2010. [Online; accessed 19-December -2021].

[SH17] Oeda Shinichi and Genki Hashimoto. Log-data clustering analysis for dropout prediction in beginner programming classes. *Procedia Computer Science*, 112:614–621, 2017. [Online; accessed 27-December-2017].

[TS15] Mingjie Tan and Peiji Shao. Prediction of student dropout in e-learning program through the use of machine learning method. *International Journal of Emerging Technologies in Learning (iJET)*, 10:11–17, 2015. [Online; accessed 7-January-2018].

[XD18] Wanli Xing and Dongping Du. Dropout prediction in moocs: Using deep learning for personalized intervention. *Journal of Educational Computing Research*, 0:23, 2018. [Online; accessed 19-December -2021].

UNIVERSIDADE DE ÉVORA
INSTITUTO DE INVESTIGAÇÃO
E FORMAÇÃO AVANÇADA

Contactos:
Universidade de Évora
Instituto de Investigação e Formação Avançada — IIFA
Palácio do Vimioso | Largo Marquês de Marialva, Apart. 94
7002 - 554 Évora | Portugal
Tel: (+351) 266 706 581
Fax: (+351) 266 744 677
email: iifa@uevora.pt