

An Analysis of Student Representation, Representative Features and Classification Algorithms to Predict Degree Dropout

Rubén Manrique
Universidad de los Andes
Bogotá, Colombia
rf.manrique@uniandes.edu.co

Bernardo Pereira Nunes
PUC-Rio / UNIRIO
Rio de Janeiro, Brazil
bnunes@inf.puc-rio.br

Olga Marino
Universidad de los Andes
Bogotá, Colombia
olmarino@uniandes.edu.co

Marco Antonio Casanova
PUC-Rio
Rio de Janeiro, Brazil
casanova@inf.puc-rio.br

Terhi Nurmikko-Fuller
Australian National University
Canberra, Australia
terhi.nurmikko-fuller@anu.edu.au

ABSTRACT

Identifying and monitoring students who are likely to dropout is a vital issue for universities. Early detection allows institutions to intervene, addressing problems and retaining students. Prior research into the early detection of at-risk students has opted for the use of predictive models, but a comprehensive assessment of the suitability of different algorithms and approaches is complicated by the large number of variable features that constitute a student's educational experience. Predictive models vary in terms of their amplitude, temporality and the learning algorithms employed. While amplitude refers to the ability of the model to operate on multiple degrees, temporality is often considered due to the natural temporal aspect of the data. In the absence of a comparative framework of learning algorithms, the aim of this paper has been to provide such an analysis, based on a proposed classification of strategies for predicting dropouts in Higher Education Institutions. Three different student representations are implemented (namely Global Feature-Based, Local Feature-Based, and Time Series) in conjunction with the appropriate learning algorithms for each of them. A description of each approach, as well as its implementation process, are presented in this paper as technical contributions. An experiment based on a dataset of student information from two degrees, namely Business Administration and Architecture, acquired through an automated management system from a university in Brazil is used. Our findings can be summarized as: (i) of the three proposed student representations, the Local Feature-Based was the most suitable approach for predicting dropout. In addition to providing high quality results, the Local Feature-Based representations are simple to build, and the construction of the model is less expensive when compared to more complex ones; (ii) as a conclusion of the results obtained via Local Feature-Based, dropout can be said to be accurately predicted using grades of a few core courses, so there is no need for a complex features extraction process; (iii) considering temporal aspects of

the data does not seem to contribute to the prediction performance although it increases computational costs as the model complexity increases.

KEYWORDS

Degree Dropout Analysis, Student Representation, Features Extraction, Temporal Analysis, Dropout Prediction

ACM Reference Format:

Rubén Manrique, Bernardo Pereira Nunes, Olga Marino, Marco Antonio Casanova, and Terhi Nurmikko-Fuller. 2019. An Analysis of Student Representation, Representative Features and Classification Algorithms to Predict Degree Dropout. In *The 9th International Learning Analytics & Knowledge Conference (LAK19), March 4–8, 2019, Tempe, AZ, USA*. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3303772.3303800>

1 INTRODUCTION

The analysis and prediction of student dropout rates in Higher Education (HE) has been the object of multiple studies in recent years [2, 7, 15, 16, 22, 23]. Decreasing the rate of dropout has become one of the main objectives for many universities in the world. In Europe, for example, the dropout rate is close to 30% according to the publication Education at a Glance (EAG) [9, 22]. Similarly in Brazil, it is estimated that only 62.4% of University enrollments succeed in getting an undergraduate degree [23]. In this discouraging context, universities and researchers have shown great interest in monitoring and predicting systems for identifying at-risk students with a risk or intention of dropping out.

Most of the research in this area has opted for supervised learning methods to build dropout prediction models [16]. However, the results are difficult to compare due to the diversity of students representations (i.e. set of features about the student extracted from data) that are used as input for the classification algorithms. As a result of the absence of a comparison of the different student representation strategies to predict dropout, there are no clear guidelines and comparison baselines on which future proposals in the area could be based. In this paper, we compare different students representations and feature extraction strategies for the dropout prediction problem. In particular, we compare student representations built to create global models (i.e. prediction models that can be applied to any degree program) and those that are focused on a particular degree program. Similarly, motivated by recent research [1] that incorporates time aspects in the representation, we consider the

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

LAK19, March 4–8, 2019, Tempe, AZ, USA

© 2019 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-6256-6/19/03...\$15.00

<https://doi.org/10.1145/3303772.3303800>

student data as a multivariate time series. The obtained results challenge some conclusions obtained in some previous works, and also provide solid indications for the type of features and most appropriate strategies for the dropout prediction problem.

This study was carried out using a dataset that contained the academic records of 2,175 students enrolled in two degrees at a Brazilian university. Our work is bounded to the use of student academic performance data and excludes additional information such as demographic information.

The paper is organized as follows: Section 2 reviews related literature on dropout prediction and time series analysis. Section 3 outlines a general approach for predicting dropout rates. Section 4 describes the set up for the experiment. The results and an in-depth discussion are presented in Section 5. Finally, the conclusions and directions for future work are discussed in Section 6.

2 RELATED WORK

Figure 1 presents a proposed taxonomy for dropout prediction strategies based on classification algorithms. The classical static classification problem¹ considers a training input-output set of m labeled examples $E = \{(x_i, y_i), i \in \mathcal{M} = \{1, 2, \dots, m\}\}$. $x_i \in \mathbb{R}^n$ is a features vector and $y_i \in \mathcal{D} = \{1, 2, \dots, d\}$ is the categorical class label associated with x_i . For the binary classification problem $d = 2$ (i.e. dropout or stay) $y_i \in \{-1, 1\}$ is assumed. The aim is to find a function $f : \mathbb{R}^n \mapsto \mathcal{D}$ from a hypothesis space of functions \mathcal{H} that minimizes a specified loss function that evaluates how well f describes the relation between input feature vectors x_i and their labels y_i . For the dropout prediction problem, we found that Naive Bayes, Logistic Regression, Random Forest, Decision Trees and K-nearest neighbors have been previously employed [7, 16, 17, 22, 23].

Regardless of the chosen algorithm, much of the performance of a classification model depends on the proper selection of the feature vector x_i . For the dropout prediction problem, the features that are used depend to a large extent on two factors: (i) the type of information that is available and (ii) the amplitude of the model that is to be designed. Some research has focused on sociodemographic characteristics of the student [7, 10], while others have focused on the academic performance records that are captured by University management systems [22]. Collecting sociodemographic data can be an expensive process due to challenges such as privacy and the truthfulness of the information provided. On the other hand, the information regarding a student’s academic performance (captured by the University progressively for each academic period) constitutes a reliable source of information. The results obtained in [7, 22] suggest that academic performance information is sufficient and in some cases more appropriate for the dropout prediction task. Our work also follows this line and only uses information about the students’ academic performances per each academic period.

The amplitude refers to the ability of the model to predict dropout in multiple degrees. Some works [17, 23] use global generic features that can be extracted from the records of any student regardless of the degree to which that student belongs. As a consequence, the resulting Global Feature-Based (GFB) prediction model can be applied to multiple degree programs. Opting to use features that

¹Hereafter the word *static* will be used to denote a difference from the temporal classification problem, which considers each input training example as a time series.

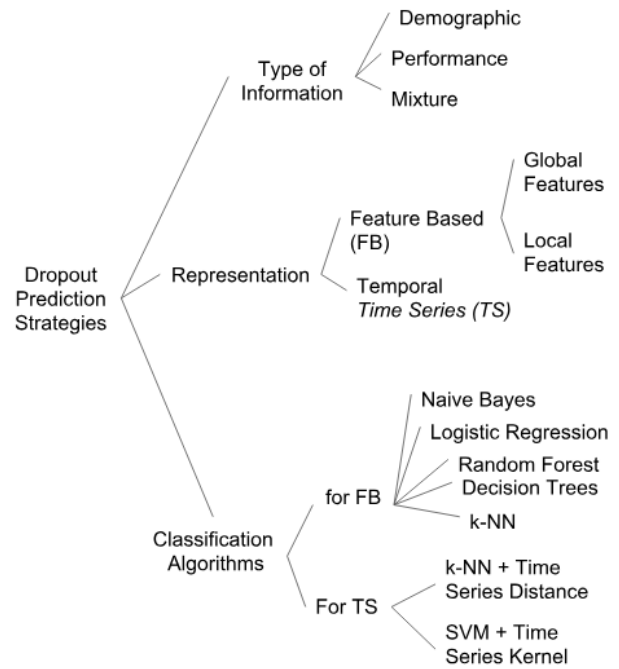


Figure 1: Classification of dropout prediction strategies

apply only to a particular degree results in a Local Feature-Based (LFB) model.

GFB models generally omit specific information from degree courses and focus on general information such as the grade averages, the number of credits enrolled and the number of lost courses [17, 23]. LFB models use as features the grades of a particular set of courses in the degree curriculum, therefore the resulting model can only be used for the degree that was considered. In [22], for example, each student in the data set is described using as a feature vector of the grades in each course enrolled in the first academic year. Similarly, [7] use the grades of 37 possible courses from the Electrical Engineering degree.

Given that the students’ records are temporal in essence, recent works have considered time series representations [1]. In the absence of or limited access to student data, it seems natural to try to exploit additional aspects such as time. Consider Figure 2, in which a student is modeled with the grades obtained in some courses (only four courses are shown for convenience). Each course is represented as a series where the sequence/time is established by the semesters. In this example the student enrolled in the course ADM1251, in the first semester and passed it with a grade of 8/10. In the third semester, the student enrolled in ADM1271 and MAT1129, but fails MAT1129 with a grade of 3/10 and passes ADM1271 with a grade of 8.7/10. In the fourth semester, the student fails MAT1129 again. Finally, in the fifth semester, he passes MAT1129 with a grade of 7.5/10. The time series encodes the number of attempts made to pass a course, the semester that the student enrolled in the course and the courses that were taken together in each academic period.

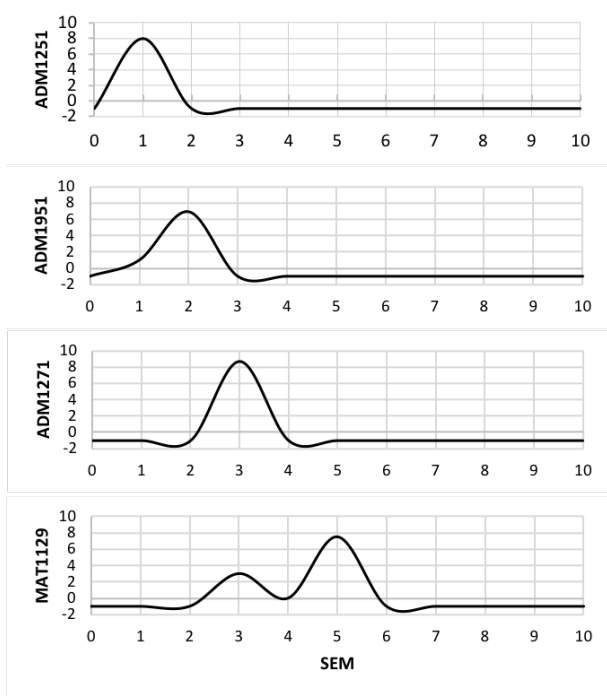


Figure 2: Smoothed time series chart for ADM1251, ADM1951, ADM1271, MAT1129 courses grades obtained by a particular student in the dataset

The hypothesis is that this additional information enriches the representation and could improve the dropout prediction models [1]. When using a representation based on time series (TS), it is however also necessary to use other types of classification strategies.

A one-dimensional (i.e. univariate) time series is a sequence of observations ordered in time (or space) where time is the independent variable [3]. Formally, a time series is defined as a sequence of real numbers $a = \{a_t \in \mathbb{R}, t \in \mathcal{T} = \{1, 2, \dots, k\}\}$, where t is a temporal index and k denotes the number of data points in a given time series (i.e. time series length). A multivariate (multi-dimensional) time series A_i is defined as a finite set of univariate time series. Therefore, in a temporal time series classification problem, the training input-output set entails a set of multivariate time series $\{A_i\}, i \in \mathcal{M}$ where each $A_i = [a_{lt}]$ is a rectangular matrix of size $u \times k$. The index associated with each univariate time series that compose the example is defined by $l \in \mathcal{L} = \{1, 2, \dots, u\}$. Here, u should be understood as the number of univariate series and the number of features in the representation. Each multivariate time series has an associated label $y_i \in \mathcal{D}$.

The objective of the multivariate time series classification problem is also to define an appropriate function f , which optimally describes the relationship between the multivariate time series $\{A_i\}$ and their labels $\{y_i\}$ [21]. Whereas in the static classification problem each example in the training dataset is a vector $x_i \in \mathbb{R}^n$, when multivariate time representations are handled there is a matrix $A_i \in \mathbb{R}^{u \times k}$. Unlike the static case, each multivariate time series

can be of a different size (i.e. different k values per example), so the native rectangular structure of the training set is not met [21].

There are two main branches for dealing with the particular characteristics of the multivariate time series classification problem. The first one transforms the time-series into a feature-based (TS-FB) representation. The second one uses a distance-based approach (TS-DB) that directly compares the distances between the time series and where classification algorithm's like k-Nearest Neighbors algorithm can be applied or a kernel method like support vector machines [8]. The latter was used in [1] to make dropout prediction.

The TS-FB approach [26] tries to reduce the time series signal to a set of statistics (such as the mean, variance, maximum value, minimum value, and entropy) that summarizes its properties. More complex characteristics such as the resulting coefficients of the Wavelet or Fourier transform are also employed as popular features [4, 20]. Using these transformations, the time series is projected into the frequency domain (Fourier transform) or the time-frequency plane (Wavelet transform). The first resulting coefficients are used as features because they correspond to the low frequencies and give a general sketch of the signal. After the feature extraction process, a representation where the static classical classification algorithms can be applied is obtained. If each univariate time series is reduced to q features, the resulting size of the feature vector for each multivariate time series will be $n = q \cdot u$. Note that the same n features will be extracted from each multivariate time series in the training set, avoiding the problem of not-equal length. In general terms, this approach makes a flattening version of the time series, and as a consequence, it is susceptible to information loss if the correct features are not chosen. Additionally, this approach can be time-consuming if the number of features extracted and/or the number of dimensions of the multivariate time series are high [26].

TS-DB strategies directly compare the raw data via different distance measures, rather than performing a feature extraction process. One of the most effective distance measures for univariate time series is the DTW (Dynamic Time Warping) distance [21]. The strategy of combining the k-neighbor classifier and DTW distance has been shown to be one of the best-performing time series classification techniques [11]. A comprehensive description of DTW can be found in [25].

3 GENERAL APPROACH TO DROPOUT PREDICTION

In this section, we summarize the different approaches used for predicting dropout. Figure 3 illustrates the general process. Each step of the process is described in further detail in what follows.

3.1 Data description and preprocessing

The process begins with unprocessed data that contains the course enrolment from semester to semester for each student, along with the grade obtained for each course. The dataset (which spans from 1996 to 2016 and was used for all the experiments), was provided by the academic system used at a Brazilian university [12]. In this paper, we used two curricula from 2002: the Architecture (ARQ) degree and the Business Administration (ADM) degree. The curricular matrices of the Brazilian courses is by credit system and all have, in their curricular matrices, elective courses, which the students are

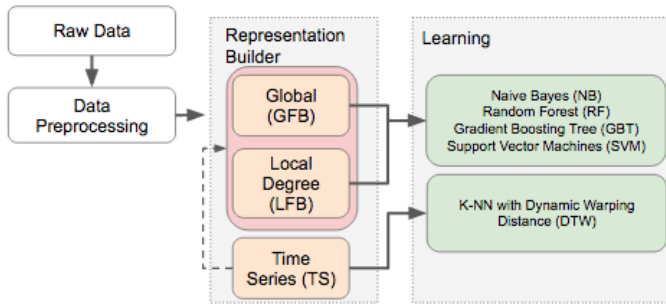


Figure 3: Dropout Prediction Process

free to take, provided they adhere to the type and number of credits established in the curriculum. The architecture degree is expected to be completed within 5 years (equivalent to 246 credits) with a maximum of 10 years. The administration degree is expected to be finished in 4 years (equivalent to 210 credits) with a maximum of 7 years. The approval criterion for the courses follows a 10-point grading. To be approved, students have to obtain an average grade of 5.0 or higher.

Each student’s record in the dataset contains information regarding their course enrollment, result status (pass or fail), final grade, course credits, effective semester, the recommended semester as well as some other details that were not included in our experiment. The *recommended semester* in this case is the semester suggested by the Brazilian university’s curricular matrix for that particular student, as the next most suitable course according to the University’s degree plan. The *effective semester* is the actual student academic semester at the time of enrollment.

Those students who registered for a semester but did not formally enroll again in the same degree were considered to have dropped out. We did not consider students who transferred to another university or HE institution as having dropped out. Similar definitions for drop outs have been used in previous works [7, 22]. Table 1 shows the number of students in each degree and the total of dropouts for each semester. For the third semester of the ADM degree, for example, there are 831 enrolled students, 100 of whom dropout in the following semesters. In the experiment, we created datasets for each semester s with the objective of predicting future dropouts. Datasets constructed for any semester s , were done so on the basis that the information for each student was also available for the semesters prior to the one considered: $1 \leq s_i < s$. This enables us to model not just the dropouts for a particular semester, but across the whole degree.

The following data cleaning operations were carried out:

- Only courses where each semester had at least 20 enrollments were considered. Courses with fewer records usually correspond to elective courses that are offered sporadically.
- When there is an enrollment record, but course grade has a missing value, we assume that the student did not complete the tasks and exams necessary to obtain a grade. These missing values were replaced by zero.
- All grades were normalized between 0 and 1. If the grade of a particular course was used to build the feature representation

Table 1: Dataset statistics per Semester (Sem) for ARQ and ADM degrees.

Sem. (s)	ARQ		ADM	
	Students enrolled in Sem. s	Dropouts after Sem. s	Students enrolled in Sem. s	Dropouts after Sem. s
1	1199	213	976	186
2	1014	129	909	144
3	907	97	831	100
4	807	64	797	82

Table 2: GFB Features

Feature	Description
#Courses	Total number of courses enrolled.
#Approved Courses	Total Number of courses approved.
#Failed Courses	Total Number of failed courses (including second attempts).
#Failed Courses (s)	Total Number of failed courses per semester s.
#Max Attempts	Maximum number of course repetitions.
Grades Mean (s)	The mean of the grades obtained per semester s.
Grades Mean	The mean of all grades obtained so far.
Diff Grades Mean	The difference between the mean grade of all students in the dataset for the considered semesters and the grades mean of the student.
Last #Courses	The number of courses enrolled in the last semester considered. Last_#Courses = #Courses only if the first semester is considered.
#Credits Completed	Total number of credits of completed courses.
#Credits Failed	Total number of credits of courses failed.
Avg Credits (s)	Average Credits per Semester s

in a semester s , but it was not enrolled in by a particular student, the value of -1 was assigned in order to differentiate it from a grade value in the representation.

3.2 Representation Builder

We built three different student representations: (i) GFB, (ii) LFB, and (iii) TS, each of which are discussed below.

3.2.1 Global Feature-based Representation (GFB). Table 2 presents the set of extracted features used to build the global feature-based representation. These types of features were employed previously by [17] and [23]. These characteristics can be extracted from the records of any degree and used to summarize the grades, credits and failed courses, without paying attention to specific degree courses.

Table 3: ADM degree. Considered courses for student representation per semester

Sem.	Considered Courses ADM degree
1	ADM1251, ADM1258, ADM1259, ADM1271, ADM1272, ADM1276, ADM1451, ADM1551, ADM1552, ADM1951, ADM1952, ADM1953, CRE1100, ECO1101, ECO1310, FIL0201, JUR1016, JUR1018, LET1040, MAT1127, MAT1128, MAT1129, PSI1033, SOC0201, SOC0203
2	ADM1251, ADM1256, ADM1258, ADM1259, ADM1271, ADM1272, ADM1275, ADM1276, ADM1277, ADM1351, ADM1451, ADM1452, ADM1453, ADM1551, ADM1552, ADM1651, ADM1951, ADM1952, ADM1953, ADM1954, CRE1100, ECO1101, ECO1310, ECO1411, FIL0201, JUR1016, JUR1018, JUR1306, LET1040, MAT1127, MAT1128, MAT1129, PSI1033, SOC0201, SOC0203
3	ADM1251, ADM1256, ADM1258, ADM1271, ADM1272, ADM1275, ADM1276, ADM1277, ADM1351, ADM1451, ADM1452, ADM1453, ADM1454, ADM1551, ADM1552, ADM1651, ADM1951, ADM1952, ADM1953, ADM1954, ADM1976, CRE1100, ECO1101, ECO1310, ECO1411, FIL0201, JUR1016, JUR1018, JUR1306, LET1040, MAT1127, MAT1128, MAT1129, PSI1033, SOC0201, SOC0203
4	ADM1251, ADM1256, ADM1258, ADM1271, ADM1272, ADM1275, ADM1276, ADM1277, ADM1351, ADM1353, ADM1451, ADM1452, ADM1453, ADM1454, ADM1551, ADM1552, ADM1651, ADM1952, ADM1953, ADM1954, ADM1973, ADM1976, CRE1100, CRE1141, ECO1101, ECO1310, ECO1411, FIL0201, HIS0201, JUR1016, JUR1018, JUR1306, LET1040, MAT1127, MAT1128, MAT1129, PSI1033, SOC0201, SOC0203

3.2.2 *Local Feature-based Representation (LFB)*. Different from the GFB approach, LFB focuses on the particular attributes of the degree. Motivated by the results obtained in [22], each student was modeled as a vector, with a consideration of the grades obtained for the courses enrolled in for each of the academic semesters. To construct this representation, all the possible courses that a student could enroll in for each academic semester of each degree were analyzed. As previously mentioned, only those courses per semester that had been enrolled by at least 20 students in the dataset were considered. Table 3 presents the set of course codes considered semester-by-semester for the ADM degree. The courses are repeated throughout the semesters. If a student failed a course in a semester and enrolled it again in the followings, we only consider the best grade obtained. In this representation, neither the number of attempts nor their associated credits were considered.

3.2.3 *Time Series Representation (TS)*. Similar to the representation proposed in [1], each student is represented as the course grades obtained throughout the semesters. In the time series, any course that the student was not enrolled in is codified with -1: after passing a course this value is maintained until the student finishes or drops out (see Figure 2).

As mentioned in Section 2, two strategies can be applied. The first directly compares the time series according to a distance measure such as DTW or GAK (we will refer to the resulting model as TS-DB). The second one extracts a set of characteristics from each time series and transforms the time series into a set of features (we will refer to the resulting model as TS-FB). In Table 4, we present the total set of features extracted for each time series. Since each student is represented by u (a time series that represents the courses that can be taken), the dimension of the resulting vector increases considerably with the increase of semesters (see Table 3). Considering only the first semester of ADM, for example, 148 features were extracted, but when considering 8 semesters, a total of 3,904 features were extracted. The resulting representation can also be categorized as a LFB because the features were extracted from particular courses for each degree.

Minimum, Maximum, Mean, Median, Sum, Quantiles, Variance, Standard Deviation, and Energy features characterize the ranges of values present in the series. LLMax, LLMin, ZVC, Skewness, Peaks, CWT coefficients, and FFT coefficients represent in time, frequency and magnitude the events and components that occurred in the series. Some of the features in Table 4 have been used in other research domains, where relevant results have been obtained [27].

3.3 Classification Algorithms

For GFB, LFB and TS-FB we employ four classical binary classifiers: Naive Bayes (NB), Support Vector Machine (SVM), Random Forest (RF), and Gradient Boosting Tree (GBT). For TS-DB, the DTW distance in conjunction with the k -NN classifier were implemented.

4 EXPERIMENTAL SETUP

Using the dataset presented in Section 3.1, our experiment compared the four prediction models described above: GFB, LFB, TS-FB and TS-DB. The classes in the datasets are balanced by oversampling the minority class. We implemented the SMOTE technique, which, instead of simply duplicating records, creates entries that are interpolations of the minority class [5]. For TS-DB, we applied a random undersampling of the majority class instead of the SMOTE algorithm, because the latter is not suitable for directly multivariate time series data. We validated the classifier models employing a 10-fold cross validation technique. Section 5 presents the average Accuracy (A), Precision (P), Recall (R) and F1-Score (F1).

The hyperparameters of each model are selected by a 10-fold cross-validated grid-search over a parameter grid. The following grid values were used (the best values found are in bold):

- (50,100,150,**200**,250,300,350) trees with a max-depth of (5,**10**,15,20,25) for RF and GBT.
- Radial basis function kernel for SVM and a regularization parameter C (0.1,**1**,10,100).
- Learning rate of (0.0001, 0.001, 0.01, **0.1**, 0.2, 0.3) for GBT.
- (1, 3, 5) were used as k values for the k -nn classifier. A value of $k=1$ is also supported by previous results that used k -nn in combination with DTW [8, 18].

There are different strategies for implementing DTW for multivariate time series. We used the independent DTW version [25].

Table 4: Features extracted for each time-series

Feature	Description
Minimum	Lowest value of the TS.
Maximum	Highest value of the TS.
Mean	Mean value of the TS.
Median	Median value of the TS.
Sum	The sum of TS values.
Quantiles	Quantile values for {10%,20%,30%,40%,50%,60%,70%,80%,90%}.
LLMax	Location of last occurrence of the maximum.
LLMin	Location of last occurrence of the minimum.
Variance	TS variance.
Standard Deviation	TS standard deviation.
Energy	Sum over the TS squared values.
ZVC	Number of zeros values in the TS.
Skewness	Skewness of TS. Skewness indicates the symmetry of the probability density function (PDF) of the amplitude of a time series.
Peaks	Calculates the number of peaks in the TS. The TS is smoothed by a Ricker wavelet and for different widths ranging from 1 to 5. The number of peaks that occur in different width scales are returned.
CWT coefficients	Coefficients of the continuous wavelet transform for the Ricker wavelet. The following standard deviation values for the wavelet function are used: (2,5,10,20).
FFT coefficients	Coefficients of the one-dimensional discrete Fourier Transform. The real, imaginary and angle values of each coefficient are returned as separate features.
Linear Trend	Calculate a linear least-squares regression for the values of the TS. The p -value, r -value, intercept, and slope are included as separate features.

5 RESULTS AND DISCUSSION

The results for the GFB representation are presented in Table 5. In general, Naive Bayes is the least suitable of all the approaches, due to the fact that the strong independence assumption does not apply for the global-based feature set. The best results were obtained with the RF and GBT ensemble models. Overall, as the number of considered semesters increases, so does the quality of the RF and GBT models. A different result was obtained in [23], who also focused on models with global features and used an RF classifier, but observed a decrease in the quality of the model. They attribute this to the fact that in later semesters, the number of dropouts decreases. In our dataset, there is also a decrease in the number of dropouts that corresponds to the increase of semesters considered,

Table 5: GFB results for both ADM and ARQ degrees combined.

		A	R	P	F1
Sem 1	NB	0.47	0.418	0.571	0.495
	SVM	0.55	0.449	0.65	0.532
	RF	0.727	0.771	0.727	0.751
	GBT	0.772	0.751	0.78	0.763
Sem 2	NB	0.478	0.429	0.532	0.482
	SVM	0.603	0.614	0.643	0.627
	RF	0.802	0.843	0.794	0.818
	GBT	0.809	0.835	0.826	0.829
Sem 3	NB	0.543	0.548	0.499	0.535
	SVM	0.608	0.621	0.588	0.604
	RF	0.832	0.843	0.825	0.835
	GBT	0.84	0.831	0.825	0.834
Sem 4	NB	0.573	0.433	0.56	0.492
	SVM	0.675	0.76	0.625	0.689
	RF	0.868	0.855	0.86	0.858
	GBT	0.826	0.861	0.823	0.842

however, we argue that with the increase of information about the student, better results should be obtained.

The results that were obtained for the LFB representation are shown in Table 6. In accordance with the GFB, the results obtained tend to improve with the increase of semesters and the best results are obtained via RF and GBT. The results also show an improvement when compared to the global model. More importantly, when using the information related to the first semester only, we can successfully predict the dropout rate with a precision higher than 80%. Similar accuracy levels were obtained by [22] using this simple representation strategy. Table 6 illustrates the greater suitability of RF in comparison to GBT with the increased number of semesters – however, this may be due to the fact that the tuning of the parameters for GBT was completed with data from the first semester.

In Figure 4, the evolution of the F1-Score with the increase of the semesters is shown for the TS-DB model. As with the previous models, the results were improved when more semesters were considered. However, the results are worse from those obtained by GFB and LFB. Intuitively, a distance measure as DTW is not the appropriate method for evaluating a series of time of such a short length. As more data points are introduced into the series, the distance measures seem to improve their capture of the notion of similarity between the representations. The Euclidean distance and the Global Alignment Kernel [6] were also implemented, but similar results were obtained. In contrast to the conclusions presented by Asknadze and Conrad [1], we conclude that a time series representation, in combination with a distance measure, seems inappropriate for the problem of predicting dropout, particularly if only data from the first academic periods are taken into consideration.

The results obtained with the TS-FB are shown in Table 7. This approach leads to results similar to those obtained by LFB (see Table 6), and substantively improves the distance-based approach (TS-DB). The similarity with LFB can partly be attributed to the fact that the features used in LFB are a subset of the features used in TS-FB. The “Maximum” feature of each time series is equivalent to the

Table 6: Results using the LFB Representation

		SEM 1				SEM 2				SEM 3				SEM 4			
		NB	SVM	RF	GBT	NB	SVM	RF	GBT	NB	SVM	RF	GBT	NB	SVM	RF	GBT
ARQ	A	0.585	0.667	0.842	0.854	0.668	0.681	0.903	0.902	0.719	0.785	0.938	0.917	0.594	0.756	0.963	0.931
	R	0.767	0.578	0.881	0.875	0.834	0.65	0.927	0.936	0.63	0.827	0.961	0.955	0.822	0.69	0.977	0.961
	P	0.584	0.764	0.835	0.864	0.645	0.733	0.892	0.883	0.773	0.787	0.922	0.893	0.568	0.8	0.953	0.909
	F1	0.652	0.643	0.852	0.86	0.721	0.677	0.907	0.906	0.692	0.798	0.94	0.921	0.67	0.74	0.964	0.933
ADM	A	0.594	0.745	0.862	0.853	0.592	0.79	0.907	0.894	0.668	0.83	0.938	0.904	0.664	0.86	0.951	0.916
	R	0.824	0.697	0.869	0.869	0.907	0.768	0.912	0.92	0.898	0.833	0.949	0.93	0.925	0.875	0.961	0.94
	P	0.566	0.776	0.86	0.844	0.558	0.808	0.907	0.88	0.618	0.834	0.931	0.885	0.61	0.856	0.944	0.899
	F1	0.67	0.733	0.861	0.853	0.69	0.786	0.904	0.894	0.731	0.831	0.939	0.905	0.734	0.862	0.95	0.916

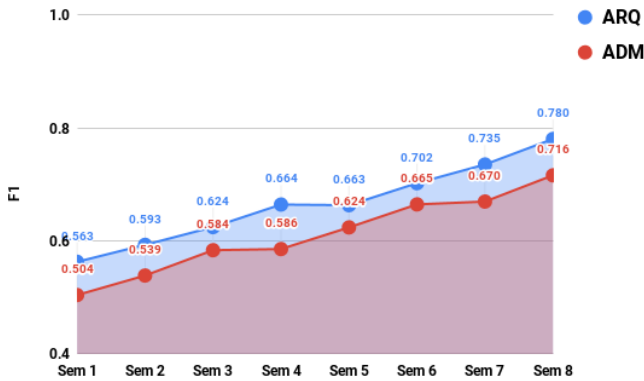


Figure 4: TS-DB Results. F1-Score per Semester is reported

maximum score obtained for each course that corresponds to the features used in LFB. However, we expected that with the increase of the semesters, the other features related to the temporal and frequency aspects would improve the prediction model. Because there is no significant improvement with the results obtained with LFB, we can say that the contribution of these additional features is not relevant, and only serve to increase the computational cost of the model.

In the experiments, we evaluated different dropout prediction models. The models differ from each other in the representation used, and as a consequence in the classification algorithms (see Figure 3). In the rest of this section, we will focus on the feasibility of the models for the problem at hand, the relevance of including temporal aspects and the implementation challenges of each representation.

The best results for the level of accuracy and the F1-Score were obtained by the LFB and TS-FB models. These models are trained with the curriculum courses of a particular degree, and cannot be generalized. In order to use the LFB and TS-BF, data from the specific degree being modelled must be used to train the model.

The main difference between LFB and TS-FB is that TS-FB uses features that are extracted from a temporal representation. Characteristics such as the number of course attempts, maximum grade, the enrollment semesters and the courses enrolled simultaneously are codified in the features extracted from the time series. LFB only

codes the highest score obtained for the set of courses considered. The results obtained by LFB (Table 6) are slightly better than the results obtained by TS-FB (Table 7). This suggests that the contribution of the temporary component is not significant for this particular dropout prediction.

This was confirmed with the results obtained by TS-DB. In this case, the time series were compared directly using DTW, which is a state-of-the-art measure of distance for the problem of time series classification [11]. Among all the models, TS-DB presented the least accurate results. We attribute the poor results to the small size of the time series, one that is not large enough to show differences between the considered examples. Figure 4 clearly shows an increase in the quality of the model with the increase of the semesters. However, since the majority of the dropouts occur in the first semesters (see Table 1), this approach should not be appropriate for the dropout prediction task. The assumption that temporal aspects are important is not properly justified and on the contrary, increases the complexity of the analysis.

It should be clarified that the time series representation may be adequate in other tasks different from the one analyzed here. Examples of such tasks are: (i) to predict dropout in a course from the analysis of the daily-generated click-stream in an educational platform [13], and (ii) to discover clusters of students with similar behaviours [19].

Regarding the GFB, we found that although it is not consistently the most suitable model, it is especially useful when there is insufficient data to build an LFB for a particular degree. To demonstrate this, we used the GFB model previously constructed from the data from the ARQ and ADM degrees to predict the dropout in the Information Systems degree. For this new degree, there were 142 students, 46 of whom we considered to be dropouts. Applying the GFB model (using GBT as a classifier) we obtained an A = 0.734 / F1 = 0.701 considering the first semester, and A = 0.773 / F1 = 0.761 for the second semester. Although the results barely exceed 70% of F1, GBF is a powerful alternative in the absence of historical data about a particular degree.

Regarding the representation and the features extraction process, we found that the TS-FB is by far the most expensive representation to build. Extracting the features for the complete dataset varies between 5min and 35min, depending on the number of semesters considered. Regarding the prediction stage, TS-DB is the most computationally costly model. The high computational cost of DTW,

Table 7: Results using the TS-FB Representation

		SEM 1				SEM 2				SEM 3				SEM 4			
		NB	SVM	RF	GBT	NB	SVM	RF	GBT	NB	SVM	RF	GBT	NB	SVM	RF	GBT
ARQ	A	0.647	0.665	0.835	0.858	0.637	0.719	0.897	0.916	0.664	0.711	0.919	0.925	0.693	0.775	0.952	0.932
	R	0.558	0.56	0.875	0.848	0.588	0.722	0.925	0.924	0.716	0.729	0.946	0.937	0.612	0.874	0.961	0.941
	P	0.738	0.767	0.827	0.878	0.692	0.751	0.887	0.921	0.678	0.707	0.908	0.92	0.738	0.735	0.95	0.933
	FI	0.621	0.634	0.844	0.847	0.625	0.729	0.902	0.916	0.688	0.717	0.923	0.922	0.667	0.796	0.951	0.928
ADM	A	0.597	0.692	0.828	0.839	0.676	0.744	0.907	0.892	0.681	0.65	0.93	0.907	0.767	0.902	0.951	0.93
	R	0.86	0.626	0.824	0.831	0.696	0.75	0.91	0.893	0.741	0.935	0.925	0.906	0.727	0.833	0.952	0.933
	P	0.567	0.729	0.833	0.847	0.679	0.747	0.906	0.893	0.667	0.596	0.934	0.908	0.797	0.966	0.95	0.929
	FI	0.682	0.672	0.826	0.836	0.685	0.744	0.902	0.882	0.701	0.728	0.923	0.9	0.758	0.894	0.947	0.926

Table 8: Top features for LFB model

		Top	Sem 1	Sem 2	Sem 3	Sem 4
ARQ	1		MAT1071	ARQ1102	ART1030	HIS1850
	2		ARQ1101	ART1029	MAT1072	MAT1072
	3		ARQ1000	ART1027	ARQ1103	FIS1011
ADM	1		ADM1251	ADM1271	ADM1551	SOC0201
	2		ADM1951	ADM1952	MAT1129	MAT1129
	3		MAT1127	MAT1128	ADM1953	ADM1258

combined with K-*nn*, was also observed in [24]. The average prediction time in our experiments² varied between 30min and 120min, depending on the number of semesters considered. One of our findings is that the most computationally efficient model is also the method that presented the best results (i.e. LFB). The feature vector for LFB was relatively straightforward to build, since minimal processing must be performed. This vector is constructed directly from the grades of a selected set of courses. All the experiments were run on a server computer with an Intel i7 CPU (4x cores) and 32 GB of RAM running Ubuntu 14.

From the results obtained for LFB, it can be concluded that the use of the course grades as feature is sufficient for the dropout prediction problem. Hence, if the course grades is a representative feature, it is possible to determine the courses that are most indicative of student dropout. For instance, using the trained RF classifiers and the LFB model (i.e. the model with the best results), we performed a feature importance analysis via the mean decrease impurity method. Table 8 lists the top three features for each domain. As LFB only uses course grades, the features correspond to the related course code.

Table 8 shows the codes of the most representative courses. The codes with the “MAT” prefix are courses in Mathematics, and those with “ARQ” and “ADM” correspond to specific courses of the discipline. An analysis with the academic administrators revealed that first semester courses were typically introductory ones. In many cases, students choose a career without extensive knowledge of the subject, and limited success in these courses is an indication of a poor selection and admission process, and an indicator of possible abandonment in coming semesters. A low grade in the Mathematics courses reveals deficiencies in the skills acquired by the students

during earlier studies (namely high school). For this reason, a failure in these courses is a suitable indicator for possible dropout. To verify this hypothesis, the distribution of students in ADM and ARQ degrees courses is presented in Figure 5. Note that the “Fail and Dropout” category is calculated as the number of students who fail the course and drop the degree in the subsequent semesters but not necessarily the immediate one. A comparison of Table 8 and Figure 5 reveals that most courses with a high “Fail and Dropout” rate correspond to courses that are more representative for the prediction. This can be interpreted as a greater probability of dropout when these particular courses are failed. Therefore, LFB shows the most accurate results, suggesting that “course performance” is a suitable predictor in our dataset.

That said, GFB models the global performance of the student without taking into account the courses as features. An analysis of false negatives revealed that GFB fails to detect dropouts of students with a “good” global performance, when students also fail in a representative course such as those previously explained and shown in Table 8. An example would be an ARQ degree student whose only failed course was ARQ1102 (a required course) in the second semester, but who scored high grades in previous courses. This suggests that GFB can be improved on by the addition of features that capture the relative importance/difficulty of the courses failed/passed, and can be automatically extracted from the data.

Finally, among the learning algorithms considered there is a clear superiority of the ensemble RF and GBT algorithms. According to Hara et al. [14] they are one of the must-try algorithms due their high prediction performance.

This analysis is however associated with a few limitations. For instance, all the experiments were carried with data representing students with a mixed curriculum. Thus, while there are electives that the student can choose in any semester, there are also courses with established prerequisites that are not necessarily available for each student to enrol in (i.e. the student can only register on a course if he has passed the prerequisite courses). Curricula in which the student is completely free to register on any and all of the courses is yet to be investigated.

Furthermore, although we consider the dataset used in our experiment as having been sufficiently large with an appropriate number of students, a more complete set of degrees should be used to repeat and verify the findings reported on here. The lack of data from other institutions also limits the experiments presented in this paper.

²The total time it took to build the results vector when applying 10 fold cross validation

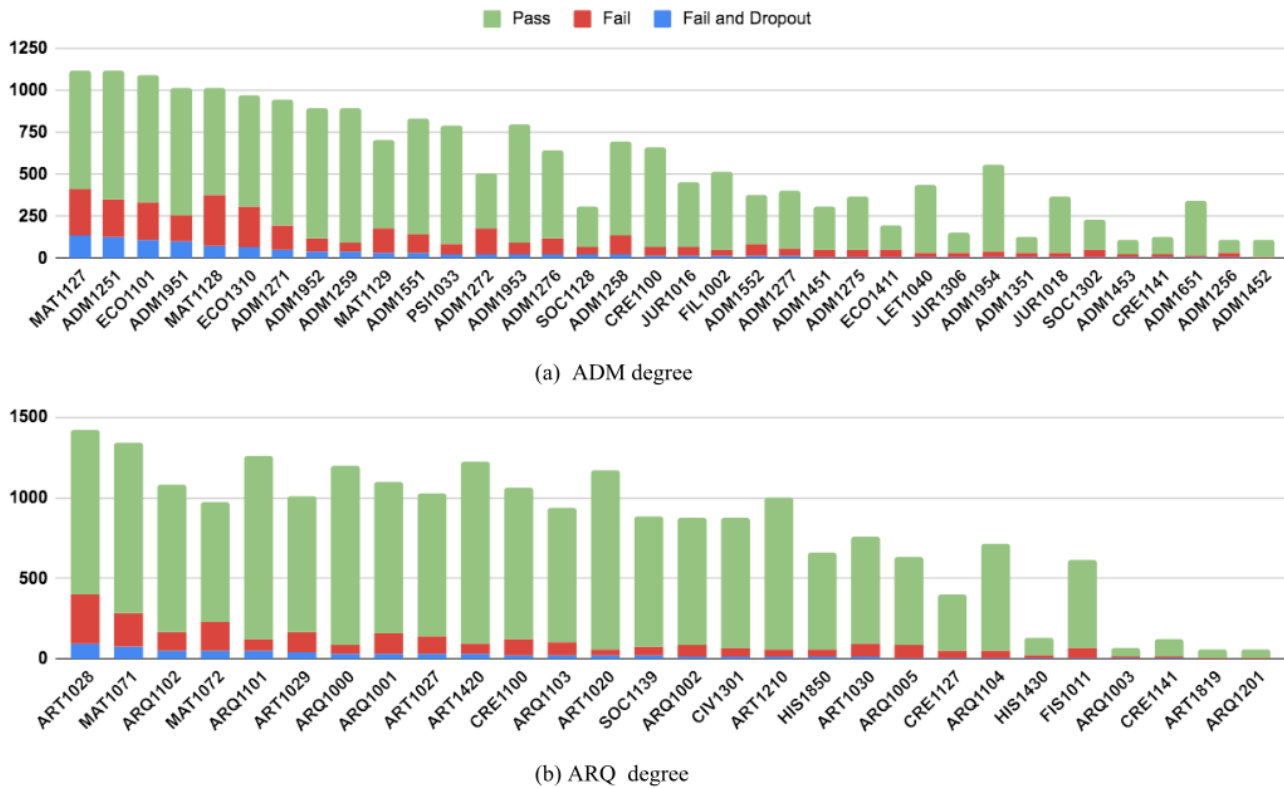


Figure 5: Student distribution per course in first 4 semesters courses. Note that only courses with more than 100/50 (ADM/ARQ) enrollments in the dataset were considered.

6 CONCLUSION

Prior research into the early detection of students at risk of dropping out from their degree programs has opted for the use of predictive models, but a comprehensive assessment of the suitability of different algorithms and approaches is complicated by the large number of variable features that constitute a student’s educational experience. The aim of this paper has been to provide a comparative analysis of available methods, based on a proposed classification of strategies for predicting dropout, and an experiment based on a dataset of student information acquired through an automated management system from a university in Brazil. Three different representations were implemented (namely GFB, LFB, and TS) in conjunction with the appropriate learning algorithms for each of them. A description of each approach, as well as its implementation process, were presented in this paper.

Finally, we summarize our three-fold findings: (i) of the three representations that were built, the LFB was the most suitable approach for predicting dropout. In addition to providing high quality results, the LFB representations are simple to build, and the construction of the model is less expensive when compared to a time series; (ii) as a conclusion of the results obtained via LFB, we can also say that dropout can be accurately predicted using grades of a few core courses, without the need for a complex features extraction process; (iii) consideration of the temporal aspects of

the data does not seem to contribute to the prediction performance, but it does increase the computational costs.

We expect that this work can serve as a direction for further investigations into student dropout rates at Higher Education Institutions. In order to do so, a more comprehensive list of degrees will be investigated as part of our future work. Similarly, we intend to run our experiments with several different datasets covering other universities’ data, gathered from different global regions and, thus, generalize the findings presented here. We also envision opportunities to include the currently omitted socio-demographic data and analyze the impact of this additional information in the predictive models.

ACKNOWLEDGMENTS

This work was partially supported by COLCIENCIAS PhD scholarship (Call 647-2014).

REFERENCES

- [1] Stefan Conrad Alexander Askinadze. 2017. Application of the Dynamic Time Warping Distance for the Student Drop-out Prediction on Time Series Data. *Proceedings of the 10th International Conference on Educational Data Mining (2017)*.
- [2] Lovenoor Aulck, Nishant Velagapudi, Joshua Blumenstock, and Jevin West. 2016. Predicting Student Dropout in Higher Education. (2016). arXiv:arXiv:1606.06364
- [3] George Edward Pelham Box and Gwilym Jenkins. 1990. *Time Series Analysis, Forecasting and Control*. Holden-Day, Inc., San Francisco, CA, USA.

- [4] Pimwadee Chaovalit, Aryya Gangopadhyay, George Karabatis, and Zhiyuan Chen. 2011. Discrete Wavelet Transform-based Time Series Analysis and Mining. *ACM Comput. Surv.* 43, 2, Article 6 (Feb. 2011), 37 pages. <https://doi.org/10.1145/1883612.1883613>
- [5] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Int. Res.* 16, 1 (June 2002), 321–357. <http://dl.acm.org/citation.cfm?id=1622407.1622416>
- [6] Marco Cuturi. 2011. Fast Global Alignment Kernels. In *ICML*, Lise Getoor and Tobias Scheffer (Eds.). Omnipress, 929–936.
- [7] Gerben Dekker, Mykola Pechenizkiy, and Jan Vleeshouwers. 2009. Predicting Students Drop Out: A Case Study. In *International Conference on Educational Data Mining*.
- [8] Hui Ding, Goce Trajcevski, Peter Scheuermann, Xiaoyue Wang, and Eamonn Keogh. 2008. Querying and Mining of Time Series Data: Experimental Comparison of Representations and Distance Measures. *Proc. VLDB Endow.* 1, 2 (Aug. 2008), 1542–1552.
- [9] The Organisation for Economic Co-operation and Development. 2013. *Education at a Glance: OECD indicators* <http://www.oecd.org/education/eag2013.htm>. Technical Report. The Organisation for Economic Co-operation and Development.
- [10] Joaquin Gairin, Xavier M. Triado, MÁsnica Feixas, Pilar Figuera, Pilar Aparicio-Chueca, and Mercedes Torrado. 2014. Student dropout rates in Catalan universities: profile and motives for disengagement. *Quality in Higher Education* 20, 2 (2014), 165–182. <https://doi.org/10.1080/13538322.2014.925230>
- [11] Tomasz Gorecki and Maciej Luczak. 2015. Multivariate time series classification with parametric derivative dynamic time warping. *Expert Systems with Applications* 42, 5 (2015), 2305 – 2312.
- [12] V. Gottin, H. JimÁñez, A. C. Finamore, M. A. Casanova, A. L. Furtado, and B. P. Nunes. 2017. An Analysis of Degree Curricula through Mining Student Records. In *2017 IEEE 17th International Conference on Advanced Learning Technologies (ICALT)*. 276–280. <https://doi.org/10.1109/ICALT.2017.54>
- [13] L. Haiyang, Z. Wang, P. Benachour, and P. Tubman. 2018. A Time Series Classification Method for Behaviour-Based Dropout Prediction. In *2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT)*. 191–195. <https://doi.org/10.1109/ICALT.2018.00052>
- [14] Satoshi Hara and Kohei Hayashi. 2018. Making Tree Ensembles Interpretable: A Bayesian Model Selection Approach. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research)*, Amos Storkey and Fernando Perez-Cruz (Eds.), Vol. 84. PMLR, Playa Blanca, Lanzarote, Canary Islands, 77–85.
- [15] Martin Hlosta, Zdenek Zdrahal, and Jaroslav Zendulka. 2017. Ouroboros: Early Identification of At-risk Students Without Models Based on Legacy Data. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference (LAK '17)*. ACM, New York, NY, USA, 6–15. <https://doi.org/10.1145/3027385.3027449>
- [16] Mukesh Kumar. 2017. Literature Survey on Educational Dropout Prediction. *International Journal of Education and Management Engineering* 7, 2 (2017).
- [17] Laci Mary B. Manhães and Geraldo Zimbrão. 2014. Evaluating Performance and Dropouts of Undergraduates Using Educational Data Mining. In *Twenty-Ninth Symposium on Applied Computing*.
- [18] Theophano Mitsa. 2010. *Temporal Data Mining* (1st ed.). Chapman & Hall/CRC.
- [19] Ewa Mlynarska, Derek Greene, and Padraig Cunningham. 2016. Time Series Clustering of Moodle Activity Data. In *AICS*.
- [20] Fabian Mörchen. 2003. Time series feature extraction for data mining using DWT and DFT.
- [21] C. Orsenigo and C. Vercellis. 2010. Combining discrete SVM and fixed cardinality warping distances for multivariate time series classification. *Pattern Recognition* 43, 11 (2010), 3787 – 3794.
- [22] Sergi Rovira, Eloi Puertas, and Laura Igual. 2017. Data-driven system to predict academic grades and dropout. *PLOS ONE* 12, 2 (02 2017), 1–21. <https://doi.org/10.1371/journal.pone.0171207>
- [23] Allan Sales, Leandro Balby, and Adalberto Cajueiro. 2016. Exploiting Academic Records for Predicting Student Drop Out: a case study in Brazilian higher education. *JIDM* 7, 2 (2016), 166–180.
- [24] Skyler Seto, Wenyu Zhang, and Yichen Zhou. 2015. Multivariate Time Series Classification Using Dynamic Time Warping Template Selection for Human Activity Recognition. *2015 IEEE Symposium Series on Computational Intelligence* (2015), 1399–1406.
- [25] Mohammad Shokoohi-Yekta, Bing Hu, Hongxia Jin, Jun Wang, and Eamonn Keogh. 2017. Generalizing DTW to the multi-dimensional case requires an adaptive approach. *Data Mining and Knowledge Discovery* 31, 1 (01 Jan 2017), 1–31.
- [26] Gian Antonio Susto, Angelo Cenedese, and Matteo Terzi. 2018. Chapter 9 - Time-Series Classification Methods: Review and Applications to Power Systems Data. In *Big Data Application in Power Systems*, Reza Arghandeh and Yuxun Zhou (Eds.). Elsevier, 179 – 220. <https://doi.org/10.1016/B978-0-12-811968-6.00009-7>
- [27] Jenna Wiens, John V. Guttag, and Eric Horvitz. 2012. Patient Risk Stratification for Hospital-associated C. Diff As a Time-series Classification Task. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1 (NIPS'12)*. Curran Associates Inc., USA, 467–475.