A comparative study of WHO and WHEN prediction approaches for early identification of university students at dropout risk

1st Daniel A. Gutierrez Pachas
Department of Computer Science
Universidad Católica San Pablo
Arequipa, Peru
dgutierrezp@ucsp.edu.pe

2nd Germain García-Zanabria
Department of Computer Science
Universidad Católica San Pablo
Arequipa, Peru
germain.garcia@ucsp.edu.pe

3rd Alex J. Cuadros-Vargas
Department of Computer Science
Universidad Católica San Pablo
Arequipa, Peru
acuadros@ucsp.edu.pe

4th Guillermo Cámara-Chávez
Computer Science Department
Federal University of Ouro Preto
Ouro Preto, Brazil
guillermo@ufop.edu.br

5th Jorge Poco
School of Applied Mathematics
Getulio Vargas Foundation
Rio de Janeiro, Brazil
jorge.poco@fgv.br

6th Erick Gomez-Nieto
Department of Computer Science
Universidad Católica San Pablo
Arequipa, Peru
emgomez@ucsp.edu.pe

***Abstract*—Reducing the students' dropout is one of the biggest challenges faced by educational institutions, especially in underdeveloped countries. Identification of the student with the highest risk of dropping out is generally used to apply corrective actions (WHO). Therefore, it is also important to determine WHEN a student will drop out, which is fundamental to planning preventive actions. In this work, we perform a study to quantitatively compare several approaches to address the early identification of dropout students in universities. We categorize our study into three main methods families, i.e., analytical methods, traditional classification methods, and probabilistic methods. The first is exploited at preprocessing step for selecting significant variables into the dropout identification task. The second uses machine learning models to classify students into dropout prone or non-dropout prone classes. The third family uses survival models to determine when the student would desert. To evaluate the predictive capacity of the classification models, the Kappa coefficient was incorporated into the usual machine learning metrics and shows that Kappa is handy for evaluating performance in unbalanced data. Similarly, in the survival models, the concordance index was applied to evaluate the predictive capacity. Our approach was applied over a real data set of Peruvian university graduate students to identify when and who will drop out.**

***Index Terms*—University dropout, Machine learning, Survival analysis.**

I. Introduction

Higher education is fundamental to build human capital, which, in turn, builds the institutions considered indispensable for a country's development [1]. Investing in education ensures a prosperous and competitive socio-economic system. In particular, higher education has a high responsibility to society because it is in charge of preparing future professionals. However, students' dropout has become one of the biggest problems that educational institutions have to face.

Student dropout is a priority problem for any educational institution in the world. This problem is very complex, where many variables and factors are involved. The high dropout rates were worrisome and little explored in the Peruvian higher education system. These rates increased much more with the health crisis caused by Covid-19. According to the Peruvian Ministry of Education (MINEDU), in 2020, the university dropout rate reached 18.6% in the whole country, six percentage points more than in 2019. Given the alarming increase in student dropouts, the Peruvian government seeks to avoid a greater number of student withdrawals by granting scholarships and distributing chips with internet access.

In [2] was formulated a theoretical model that explains the processes of interaction between the individual and the university that leads differing individuals to drop out from institutions of higher education, and that also distinguishes between those processes that result from indefinably different forms of dropout behavior. In addition, explore various definitions of student dropout.

For analyzing dropout causes, the literature has identified many factors. According to many works, academic performance is the main cause to be considered for dropout analysis [3]. However, this variable is not decisive in identifying students at risk of dropping out. For instance, socioeconomic factors also could be determinant to predispose a dropout [3, 4]. In summary, multiple factors can be categorized as student-related, family-related, and school-related variables [5]. Moreover, it is also important to consider temporal information about these factors [6]. It is clear that the students' dropout depends on several factors, and any analysis in this context must be able to deal with multiple factors and temporal information. Due to its wide scope, this topic can be formulated through many different perspectives, allowing

for a wide variety of analyzes. From this problem arose two big questions: **(1) What are the factors that cause student dropouts?**, **(2) How long does it take for a student to drop out?**.

To answer the first question, the application of traditional machine learning models has been used. Some researches address the problem of student dropout in Latin American institutions using traditional classification methods of machine learning [7–9]. Recently, [9] presented a case study in a private Peruvian university where compare (in terms of accuracy) Bayesian network techniques with decision trees. In respect to the second question, survival analysis has been used. Survival analysis refers to a branch of statistical analysis that evaluates the effect of predictors on time until an event, rather than the probability of an event, occurs. In this context, some research questions are: **What is the impact of certain characteristics on student dropout? What is the probability that a student survives 2 years? Are there differences in survival between groups of students?**. The usual metric to respond to these questions is given by the estimation of survival probability function and hazard function. Some case studies by this approach, we found in [6, 10].

This paper seeks to integrate these two approaches using the demographic, academic, and temporal information of students from the San Pablo Catholic University (UCSP) in Peru. For the first approach, machine learning models such as Logistic Regression, Support Vector Machine, Naive Bayes, Decision Tree, and Random Forest will be applied. These models will be compared based on the usual metrics as Accuracy, Precision, Recall, F1, and the area under the curve ROC (or simply AUC). Also, the predictive capacity of these models is evaluated with the Kappa coefficient.

On the other hand, to answer the second question, survival models will be applied. In the first instance, we explore the survival univariate models. In these cases, the survival is according to one feature under investigation but ignores the impact of any others. Based on these models, we estimate the survival probability and cumulative hazard functions, using the Kaplan Meier estimator and the Nelson Aelen estimator, respectively. In addition, a comparative study is based on three points of view (graphical, analytic, and by hypothesis contrast) to show if gender's student is a significant variable or not. Finally, to evaluate the influence of the covariates on the level of risk of survival, we used the Cox proportional-hazards model.

In summary, the main contributions of this work are:

- An exhaustive comparative analysis machine learning classification models and survival analysis methods to determine which one provides the highest predictive power.
- Deep analysis of features to identify their influence on the level of risk of student dropout with real data from a Peruvian university.
- A data-driven methodology to answer Who students and When probably would drop out of the university.

II. Related Work

The literature about student dropout is extensive, so to contextualize our proposal, we grouped our review into three groups: analytical approaches, traditional classification models, and probabilistic models.

A. Analytical Approaches

Studies in this area have taken a purely statistical approach to predict a student dropout. These investigations generally collect, filter, and select data to perform a correlation analysis between these characteristics and the student dropout label. In addition to correlation studies, the statistical distributions of selected characteristics are explored. We must note that this approach is descriptive and not predictive of future dropouts. Its main use is to understand better the data and subsequent communication of the results to design adequate prediction methodologies. In [11] was examined, dropout as a measure of school performance and compare according to urban or suburban origin. This study explored the distributions of dropout and turnover rates among many United States high schools and tested a series of models to explain these differences.

In another context, [12] seeks to accurately predict student performance and provide a means of identifying struggling students. It also recommends intervention strategies for at-risk students (those with the highest probability of dropping out) and helping them remain. This work also analyzes the relationship between these students' characteristics and groups them according to academic, psychological, sociological, and external factors. On the other hand, [13] applied the expectancy theory of motivation to predict the academic performance of male students and obtained that personality variables of self-esteem, internal-external control, and dogmatism all moderated the relationship between expectancy beliefs and effort.

Also, [14] conducted a study that seeks to understand the high dropout rates, especially in science, and link them to the lack of basic skills in students entering university and [15] focused on the education of distance learning and aims to investigate the main causes for student dropouts.

B. Traditional Classification Methods

The majority of works address the student dropout problem by relying on machine learning algorithms. In most of these works, the dropout rate is defined as the number of students who register for a course and they did not formally enroll again for the next two consecutive academic years. [16] proposed a data-driven system to extract relevant information hidden in the student academic data based on machine learning techniques. Additionally, presented different visualizations which help in the interpretation of the results.

Also, [17] presented an academic analytic investigation into the modeling of academic performance of engineering students enrolled in a second-year class. The modeling method used was binary logistic regression. The target predicted variable was "*success status*" defined as those students from the total originally enrolled group that achieved a final unit grade of pass or better. In the same vein, [18] applied genetic algorithms

to select a subset of artificial neural networks and functions to predict high-risk students leaving school in the first year at Virginia Commonwealth University. Other methodologies include classification algorithms, including Naive-Bayes classifier [19], Support Vector Machine [20], K-Nearest Neighbor [21] Random Forest [22] and Decision Tree [9, 23]. These models consistently outperformed rule-based models on traditional metrics such as precision, recall, and AUC. Besides, models such as Bayesian networks were employed to identify students who were likely to fail in mathematics courses [24]. Also, [25] designed several evaluation metrics to assess the goodness of machine learning algorithms from an educator's perspective. In [26] was formulated a visual analytic tool for analyzing student admission denominated *PerformanceVis*.

In the context of Online Courses, various works summarize the application of machine learning and data mining techniques, as we see in [27]. Some contributions of Prenkaj et al. [27] are a comprehensive hierarchical classification of existing literature that follows the workflow of design choices in the student's dropout facilitate the comparative analysis, associated with alternative dropout models and an exhaustive revision of machine learning methods recently proposed.

C. Probabilistic Methods

Although the classification methods predict attrition, they do not consider the temporal evolution of the attrition rates, as they do consider survival analysis investigations. Briefly, Survival Analysis involves considering the time between a fixed starting point (for example, the student's first semester in college) and a final event (for example, the student's dropout). For these types of models, the event (that is, the student's dropout) will not necessarily have occurred in all students.

In [6] was developed a Cox regression model (time-dependent) using the pre-enrollment data and their semi-annual information. This model was built around a statistical survival model and estimated the students who will continue their studies. However, this approach does not consider correlation in the data after first attrition occurs, as [28] did.

In another context, [29], [30], and [31] defined an approach based on Markov chains. Additionally, [32] developed an instrument to measure the latent trait propensity to drop out in face-to-face higher education based on item's response theory. This theory refers to a family of latent trait models used to establish psychometric properties of the items and scales. An integration of this theory with the classical models of machine learning can be found in [33].

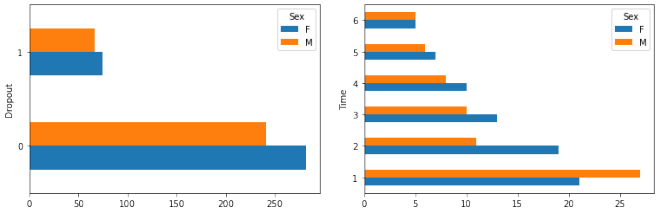
Note: Despite the analytical approaches, traditional classification and probabilistic methods are useful for analyzing different phenomena; they work independently in different contexts. For instance, analytical approaches and traditional classification methods are worried about identifying/predict WHO will drop out. On the other hand, probabilistic methods, widely used in the Medical Context, are worried about determining the time of occurrence of a phenomenon. In contrast to these methods described, our approach combines machine learning with probabilistic methods to determine who will drop out and determine when this student will drop out. Moreover, it is not necessarily split the data for each semester to evaluate the predictive accuracy in this work.

III. Our Approach

This section presents the dataset and briefly describes the approaches used.

A. Dataset

For our study, we considered a dataset with 665 students (divided into 357 female students and 308 male students) of the UCSP from the first semester of 2012 to the second semester of 2015 (8 academic semesters). This dataset contains demographic and academic variables defined into 21 features, including a binary variable that defines the student's *dropout*. TABLE I summarizes these variables.



(a) Number of students. (b) Number of dropout students.

Fig. 1. Number of students at UCSP grouped by sex and time.

TABLE I
DESCRIPTION OF THE DATA ATTRIBUTES COLLECTED FROM EACH STUDENT AT UCSP.

Nº	Variable
1	Student code
2	Number of semesters
3	Student status by semester
4	Date of Birth
5	Sex
6	Number of courses in the semester
7	Semester average
8	Standard deviation of the semester average
9	Number of absences in the semester
10	Absence rate by course
11	Average number of students per course
12	Current semester
13	Number of semesters studied
14	Semester that the desertion occurred
15	Number of accumulated absences
16	Number of accumulated courses
17	College entrance semester
18	Number of courses failed in the semester
19	Difference from student average
20	Student semester average trend
21	Student Dropout

Exploring the dataset, we found that approximately 85% of students who dropped out occurred during the first two years (four semesters). In addition, there were no dropout students in times seven and eight, which correspond to the fourth year of study. In Fig. 1(a), we illustrate the number of students grouped by gender and if they dropped out or not.

Moreover, Fig. 1(b) shows the number of dropout students classified according to time. Notice that most of the students' dropouts occurred in the first semester. Moreover, after the sixth semester, there are no dropout students.

B. Methods

The purpose of this study is to predict students **WHO** dropout and **WHEN** they drop out. Based on the features of the dataset given in TABLE I. We consider an exploratory analysis of these attributes in order to obtain the variables or the ratio between them. As the first step was analyzed, the correlation of the features to identify independent variables. Once we selected independent variables, we have interested to add temporal information. However, it is not a trivial task because there is a predominant temporal effect, and it could not guarantee the consistency of the values. To preserve the consistency of the variables, new variables are defined based on the original values.

In addition, these variables were selected based on the opinion of education experts and adapted for the application of Machine learning classification algorithms and Survival analysis methods.

Finally, we consider the following features:

- T : Number of semesters.
- X_1 : Student's gender.
- X_2 : Number of absences per course.
- X_3 : Number of failures per course.
- X_4 : Number of courses per semester.
- X_5 : Student performance.
- Y : Student's dropout.

We write T to represent the vector whose components are the number of a semesters. In addition, the vector of covariates $X = (X_1, \ldots, X_5)$ is composed by time-invariant vectors. Based on the variables defined in Table I, we have that X_1 is preserved in the time, and the other components of X are defined as follows:

$$X_2 = \frac{\text{Number of accumulated absences}}{\text{Number of accumulated courses}},$$

$$X_3 = \frac{\text{Number of accumulated failures}}{\text{Number of accumulated courses}},$$

$$X_4 = \frac{\text{Number of accumulated courses}}{\text{Number of semesters}}.$$

Also, X_5 represent the GPA (Grade Point Average) of each student. In our approaches, we use the same predictor variables, defined by $X = (X_1, \ldots, X_5)$.

However, the target variables are different. As usual in the machine learning classification algorithms (*Approach 1*), Y allows us to implement machine learning classification models to predict which students dropped out. Differently in the survival analysis methods (*Approach 2*), the target variable is defined by the pair (Y,T), where T represents the time in which the student drops out, otherwise (for censored data) is the last semester that the student was enrolled.

Approach 1: Machine learning classification algorithms

We seek to answer the question about **WHO** will drop out. Given a machine learning model:

$$\mathcal{F}(X) = Y. \tag{1}$$

Our interest is to determine the predictive capacity of the variables and choose the best predictive model. The literature on these methods is extensive, and their applications are very diverse, and they even address the problem of student dropout. For example, if the Equation (1) represent a logistic regression model, as

$$\log(\text{Odds}(X)) = \theta_0 + \theta_1 X_1 + \cdots + \theta_5 X_5. \tag{2}$$

Given an index $\ell \in \{1, \ldots, 5\}$ we calculate ODDS_ℓ as follows:

$$\text{ODDS}_\ell = \frac{\text{ODDS}(X_1, \ldots, X_{\ell+1}, \ldots, X_5)}{\text{ODDS}(X_1, \ldots, X_\ell, \ldots, X_5)} = \exp(\theta_\ell). \tag{3}$$

The methods used in this work are Logistic Regression (LR), Support Vector Machine (SVM), Naive Bayes (NB), Decision Trees (DT), and Random Forest (RF). For our experiments, we we take 70% of the sample to estimate the algorithm (denominated X_{train}) with specific parameters and 30% (denominated X_{test}) to classify the observations. This procedure is replicated five times. Finally, the classification results of these five replications are averaged.

The predictive capacity of the algorithm is evaluated with usual metrics such as Accuracy, Precision, Recall, F1, and the area under the curve ROC (or simply AUC). This is replicated five times. At the end of the process, the classification results of these five replications are averaged. Additional to well-known metrics (Accuracy, Precision, Recall, F1, and AUC), we used the Kappa statistic (or simply Kappa). Kappa adjust accuracy by accounting for the possibility of a correct prediction by chance alone. This metric is computed as follows:

$$\text{Kappa} = \frac{p_0 - p_e}{1 - p_e} = 1 - \frac{1 - p_0}{1 - p_e}, \tag{4}$$

where p_o is the relative observed agreement among raters, and p_e is the hypothetical probability of chance agreement. Depending on the model used in Equation (1), the interpretation of Kappa, defined in Equation (4), is variant.

For example, [34] considers a good performance of the model for kappa values greater than 0.6. Kappa is a very useful but under-utilized metric and can be used when measures such as accuracy, precision, or recall do not provide the complete picture of the performance of our classifier. In some other cases, we might face a problem with imbalanced classes. For instance, if we have two classes, say A and B, and A shows up on 5% of the time. Accuracy can be misleading, so we go for measures such as precision and recall. There are ways to combine the two, such as the F1, but this metric does not have a very good intuitive explanation. Then, Kappa is a very good measure that can handle very well imbalanced class problems.

Approach 2: Survival analysis methods

We seek to answer how long a student stays (**WHEN**). Survival analysis is a set of probabilistic models that will help us answer this question. These methodologies consist of statistical methods for longitudinal data analysis on the occurrence of events. Survival analysis methods are different from typical regression/classification because it depends on T. It's also possible that the student never dropout, so we won't know if the student dropout or not. Hence, for the ℓ-th student, we conclude that T_ℓ is either:

- Actual time-to-dropout if we get to observe it.
- Last time, we know that the student has not dropout. To know which of the two cases happened, we consider the variable $Y_\ell = 1$ if we got to see the student dropout, $Y_\ell = 0$ otherwise. If $Y_\ell = 0$, we said that ℓ-th time student is said to be censored.

Nevertheless, an important distinction among modeling methods is the type of outcome variable being used [35]. In survival analysis, the outcome variable is "*time to an event*", and there may be censored data. In linear regression modeling, the outcome variable is generally continuous, and in logistic modeling, the outcome variable is dichotomous (yes or not). The survival probability function is defined by:

$$S(t) = \text{Prob}(T > t). \tag{5}$$

which represent the probability that a student's dropout has not occurred yet at time t. In addition, the hazard function given by:

$$h(t) = \lim_{\delta t \to 0} \frac{\text{Prob}(t \leq T \leq t + \delta t \,|\, t \leq T)}{\delta t}, \tag{6}$$

which computes the probability of a student's dropout occurring at time t. In addition, $H(t)$ represent the cumulative hazard function. The relationship between $S(t)$ and $h(t)$ can be expressed equivalently by:

$$h(t) = -\frac{S'(t)}{S(t)}.$$

In the literature of survival analysis, we have several methods to estimate (5) and (6). One of them is a non-parametric estimator, such as Kaplan Meier estimator (KM-estimator) and Nelson Aelen estimator (NA-estimator). In this paper, we applied KM-estimator to estimate $S(t)$ and NA-estimator to approximate $H(t)$, which are denoted by $\hat{S}(t)$ and $\hat{H}(t)$ respectively. In contrast with survival models defined above. One of the popular methods in survival analysis is the Cox proportional hazard model (or simply, the Cox model). The Cox model provides a useful and easy way to interpret information regarding the relationship of the hazard function. The hazard function for the Cox model can be written as:

$$\log(h(t|X)) = \log(h_0(t)) + \theta_1 X_1 + \cdots + \theta_5 X_5, \tag{7}$$

where $h_0(t)$ is in the baseline hazard function. Also, $S(t|X)$ represent the survival probability function conditioned to predict variables X. A survival analysis aims to obtain some measure of the effect that describes the exposure-outcome relationship, adjusted for relevant extraneous variables. In logistic regression, see Equation (2), the measure of effect is given by the odds ratio, defined in (3). Similarly, in the Cox regression, see Equation (7), the measure of effect typically obtained is called a hazard ratio. Given an index, $\ell \in \{1, \ldots, 5\}$ we calculate the hazard ratio of the ℓ-th variable, denoted by HR_ℓ, as follows:

$$\text{HR}_\ell = \frac{h(t|(X_1, \ldots, X_{\ell+1}, \ldots, X_5))}{h(t|(X_1, \ldots, X_\ell, \ldots, X_5))} = \exp(\theta_\ell). \tag{8}$$

Hazard ratios are measures of association widely used in prospective studies. It is the result of comparing the hazard function among exposed to the hazard function among non-exposed. As for the other measures of association, $\text{HR}_\ell = 1$ means lack of association, $\text{HR}_\ell > 1$ suggests an increased risk, and $\text{HR}_\ell < 1$ suggests a smaller risk. In another context, a censoring-sensitive measure is the concordance index (or simply C-index). Based on [36], we compute the C-index in the following way: For every pair of students i and j (with $i \neq j$), look at their risk scores (η_i and η_j) and times-to-event (T_i and T_j) we have:

- If both T_i and T_j are not censored, then we can observe when both students got the dropout. Then (i, j) is a concordant pair if $\eta_i > \eta_j$ and $T_i < T_j$, and it is a discordant pair if $\eta_i > \eta_j$ and $T_i > T_j$.
- If both T_i and T_j are censored, then we don't know who got the dropout first (if at all), so we don't consider this pair in the computation.
- If one of T_i and T_j is censored, we only observe one dropout. Let's say we observe student i getting disease at time T_i, and that T_j is censored. (The same logic holds for the reverse situation.)
  - If $T_j < T_i$, then we don't know for sure who got the dropout first, so we don't consider this pair in the computation.
  - If $T_j > T_i$, then we know for sure that student i got the drop out first. Hence, (i, j) is a concordant pair if $\eta_i > \eta_j$, and is a discordant pair if $\eta_i < \eta_j$.

Finally,

$$\text{C-index} = \frac{\#\,\text{concordant pairs}}{\#\,\text{concordant pairs} + \#\,\text{discordant pairs}}. \tag{9}$$

This measure evaluates the accuracy of the ranking of the predicted time. In addition, (9) is a generalization of usual metric AUC. C-index can be understood as follows: 0.5 is the expected result from random predictions, 1.0 is a perfect concordance, and 0.0 is perfect anti-concordance.

IV. Results

In this section, we present the results obtained by the two approaches presented in this paper. The first approach is to predict students **WHO** dropout using Machine Learning Classification Algorithms, and the other approach is to predict **WHEN** the student will drop out.

Approach 1: Machine Learning Classification Models

As usual in the literature, we consider some metrics to evaluate the performance predict of Logistic Regression (LR), Support Vector Machine (SVM), Naive-Bayes (NB), Decision Tree (DT), and Random Forest (RF). The metrics selected are Accuracy, Precision, Recall, F1, AUC, and Kappa coefficient. For each method, we have computed five experiments. Each procedure takes 70% of the sample to estimate the algorithm with specific parameters and 30% to classify the observations.

Table II, III, IV, V, and VI shows the results of the metrics for LR, SVM, NB, DT, and RF, respectively. Table VII shows the average results for each method and metric. As we can see, RF presents the best values for almost all metrics. RF only gets a lower Recall value with DT. However, DT presents the worst performance as a function of AUC and Kappa.

TABLE II
METRICS OBTAINED WITH LR IN FIVE ALEATORY EXPERIMENTS.

#	Accuracy	Precision	Recall	F1	AUC	Kappa
1	0.698	0.743	0.605	0.667	0.804	0.395
2	0.732	0.727	0.744	0.736	0.809	0.465
3	0.744	0.811	0.667	0.733	0.803	0.493
4	0.721	0.706	0.632	0.667	0.780	0.428
5	0.779	0.846	0.717	0.776	0.867	0.561
Mean	**0.738**	**0.773**	**0.697**	**0.730**	**0.828**	**0.477**

TABLE III
METRICS OBTAINED WITH SVM IN FIVE ALEATORY EXPERIMENTS.

#	Accuracy	Precision	Recall	F1	AUC	Kappa
1	0.756	0.844	0.628	0.720	0.825	0.512
2	0.767	0.795	0.721	0.7566	0.812	0.535
3	0.756	0.900	0.600	0.720	0.814	0.518
4	0.733	0.727	0.632	0.676	0.797	0.450
5	0.802	0.968	0.652	0.779	0.879	0.612
Mean	**0.763**	**0.846**	**0.668**	**0.742**	**0.841**	**0.529**

TABLE IV
METRICS OBTAINED WITH NB IN FIVE ALEATORY EXPERIMENTS.

#	Accuracy	Precision	Recall	F1	AUC	Kappa
1	0.721	0.788	0.605	0.684	0.822	0.441
2	0.756	0.824	0.651	0.7276	0.808	0.512
3	0.733	0.867	0.578	0.693	0.805	0.473
4	0.721	0.750	0.552	0.636	0.788	0.418
5	0.779	0.909	0.652	0.759	0.851	0.565
Mean	**0.755**	**0.845**	**0.637**	**0.725**	**0.833**	**0.510**

TABLE V
METRICS OBTAINED WITH DT IN FIVE ALEATORY EXPERIMENTS.

#	Accuracy	Precision	Recall	F1	AUC	Kappa
1	0.756	0.775	0.721	0.747	0.756	0.512
2	0.698	0.730	0.628	0.675	0.698	0.395
3	0.756	0.800	0.711	0.753	0.758	0.513
4	0.686	0.628	0.711	0.667	0.699	0.372
5	0.826	0.860	0.804	0.831	0.827	0.651
Mean	**0.737**	**0.749**	**0.742**	**0.742**	**0.738**	**0.473**

As commented before, Kappa defined in (4), is a very useful metric. It gives us a better interpretation of the results when the usual metrics present problems such as unbalance in their predictions. In this vein, only RF has a value over 0.60, reinforcing that this algorithm has the best performance.

TABLE VI
METRICS OBTAINED WITH RF IN FIVE ALEATORY EXPERIMENTS.

#	Accuracy	Precision	Recall	F1	AUC	Kappa
1	0.767	0.829	0.674	0.744	0.865	0.535
2	0.767	0.829	0.674	0.744	0.812	0.535
3	0.744	0.926	0.555	0.694	0.853	0.497
4	0.744	0.735	0.657	0.694	0.827	0.476
5	0.872	1	0.760	0.864	0.905	0.747
Mean	**0.785**	**0.857**	**0.702**	**0.768**	**0.870**	**0.570**

TABLE VII
AVERAGE OF METRICS FOR EACH METHOD.

Method	Accuracy	Precision	Recall	F1	AUC	Kappa
LR	0.738	0.773	0.697	0.730	0.828	0.477
SVM	0.763	0.846	0.668	0.742	0.841	0.529
NB	0.755	0.845	0.637	0.725	0.833	0.510
DT	0.737	0.749	**0.742**	0.742	0.738	0.473
RF	**0.785**	**0.857**	0.702	**0.768**	**0.870**	**0.570**

Approach 2: Survival analysis methods

For this approach, we use the Kaplan Meier estimator to determine the estimated survivor curves for all students and students grouped by gender. We write:

- $S(t)$: Survival probability function of students.
- $S_1(t)$: Survival probability function of male students.
- $S_2(t)$: Survival probability function of female students.

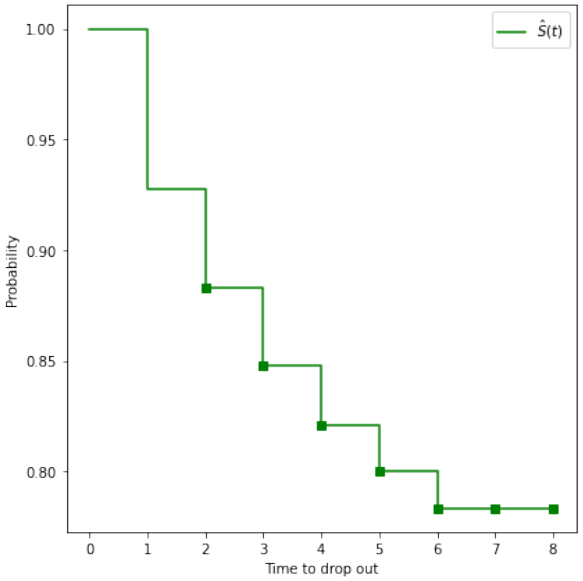


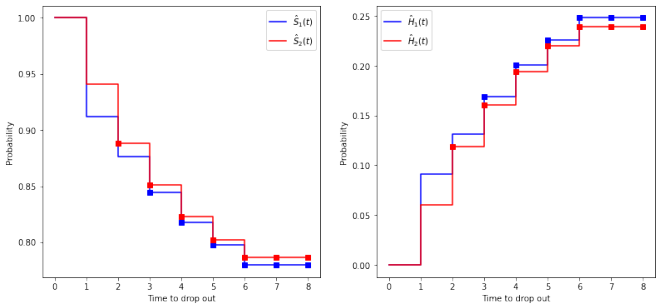
Fig. 2. Estimation of survivor curve for all students with censored data.

Also, we denote by $H(t)$ to define the cumulative hazard function. Analogously, $H_1(t)$ represents the cumulative hazard function of male students, and $H_2(t)$ represents the cumulative hazard function of female students.

TABLE VIII
ESTIMATION OF SURVIVAL PROBABILITY FUNCTIONS BY ALL STUDENTS AND SUBGROUPS OF MALE AND FEMALE STUDENTS.

t	1	2	3	4	5	6	7	8
$\hat{S}(t)$	0.93	0.88	0.85	0.82	0.80	0.78	0.78	0.78
$\hat{S}_1(t)$	0.91	0.88	0.84	0.82	0.80	0.78	0.78	0.78
$\hat{S}_2(t)$	0.94	0.89	0.85	0.82	0.80	0.79	0.79	0.79

The NA-estimator computes these curves. Graphically, we note that both curves are very similar, see Figure 3. Notice that censored data appears from the second semester for the entire group. Besides, we compare the survivor curves for the subgroups of students composed of male and female students in Fig. 3(a) and Fig. 3(b), respectively.



(a) Survival probability function. (b) Cumulative hazard function.

Fig. 3. Estimation of the Survival probability function and the Cumulative hazard function using the KM-estimator and the NA-estimator, respectively.

On the other hand, using the Log-rank test, whose null hypothesis is $\mathcal{H}_0 : h_1(t) = h_2(t)$ and obtain a p-value=0.8. That means that there is insufficient evidence to reject the null hypothesis, and consequently, the groups of male and female students are similar. That means that there is insufficient evidence to reject the null hypothesis, and consequently, the groups of male and female students are similar. Finally, we show through three points of view (graphical, analytical, and by hypothesis contrast) that the student's gender does not influence the risk of dropping out.

Now, we estimate the hazard function for the Cox model. Assuming L_1-ratio equal to 1 and a penalize equal to 0.0005. We summarize the parameters θ_ℓ, $\ell = 1, \ldots, 5$ in Table IX.

TABLE IX
SUMMARY OF PARAMETERS θ_ℓ, $\ell = 1, \ldots, 5$ COMPUTED BY COX MODEL.

ℓ	θ_ℓ	HR_ℓ	θ_ℓ lower 95%	θ_ℓ upper 95%	p-value
1	-0.10	0.90	-0.44	0.23	0.55
2	-0.01	0.99	-0.05	0.03	0.54
3	4.46	**86.26**	2.71	6.20	< 0.005
4	-0.22	**0.80**	-0.38	-0.06	0.01
5	0.02	1.02	-0.20	0.23	0.89

According to Table IX we show that X_3 (failures per semester) and X_4 (courses per semester) are more significant. These results reinforce the idea that those students who fail the most are those most at risk of dropping out. While those students who take more courses per semester represent those, who are more persevering, their risk of dropping out is lower.

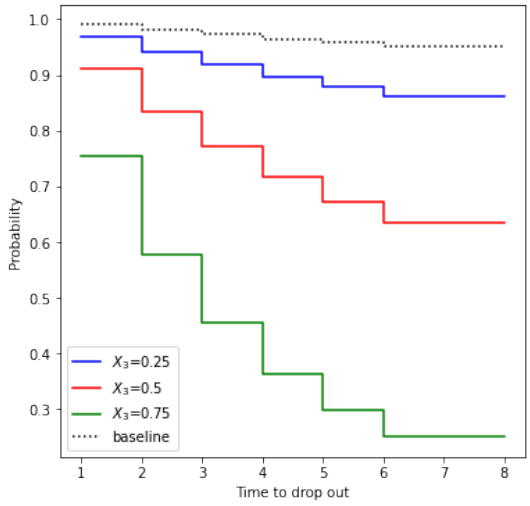


Fig. 4. Representation of survival curves varying X_3.

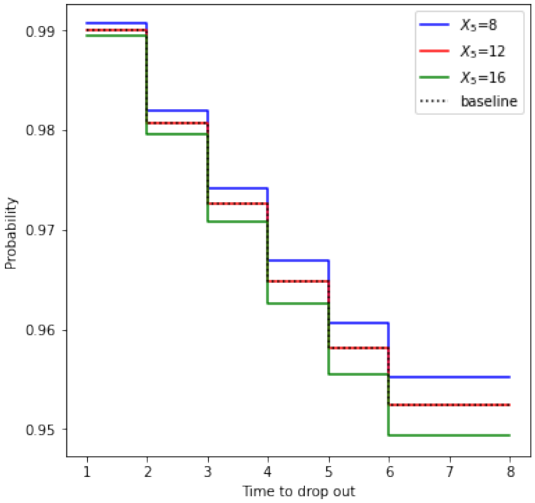


Fig. 5. Representation of survival curves varying X_5.

Fig. 4 shows the effect of the variable X_3 that is very high, $X_3 = 0.25$ means that the student fails one of four courses. We also note that as this value increases we note that the probability of survival decreases abruptly. In contrast, the variable X_5 does not represent a significant effect, which is illustrated in Fig. 5. Also, depending on the hazard ratio, defined in (8), we note that $\text{HR}_3 = 86.26$, and this means that the exposed group has 86.26 times the hazard of the unexposed group. Differently, $\text{HR}_4 = 0.80$ implies that the exposed group has eight-tenth the hazard of the unexposed group. The other variables (X_1, X_2 and X_5) have HR_ℓ, $\ell = 1, 2, 5$ close to one. That means that X_1 (Student's gender), X_2 (Absences per course), and X_5 (Student performance) do not represent a predominant effect on student dropout risk.

Even though the variable corresponding to GPA in other papers is a predominant factor, predicting the student will drop out. In this paper, we found that this variable is not a relevant factor. In the pre-processing, we identified students with good GPA drop out in the first two semesters. Then we can conclude that other causes produce the dropout of these students. Similar to machine learning classification algorithms, we show the predicted capacity of the Cox model in terms of C-index, which is defined in Equation (9). For our computational implementation, we divide the data set into training and testing data according to the first 70% of sample data as X_{train} and the rest as X_{test}. We performed five randomized experiments with the obtained samples and obtain that all the experiments exceed 70% of the C-index except for the second experiment. Therefore the mean of the experiments a value greater than 70% is maintained. Therefore Cox model predicts a student's dropout time very well.

Illustrative Example

To illustrate our proposal, we select a sample of five students with the characteristics described in the Table X. Also, we consider Logistic Regression (*Approach 1*) and Cox Regression (*Approach 2*) to illustrate our proposal.

TABLE X
FEATURES OF RANDOM SAMPLE OF FIVE STUDENTS AT UCSP.

Student	X_1	X_2	X_3	X_4	X_5
1	F	8.67	0	6	12.99
2	M	24.40	1	5	6.45
3	F	11.15	0.15	6.5	13.24
4	M	14.69	0.29	5.25	12.02
5	F	21.20	0.48	6.25	9.93

Logistic regression seeks to determine those who dropout. With the sample data with features given in the Table XI and we predict that Student 2, Student 4, and Student 5 will drop out. As can see in Table X we obtain that Student 1 and student 3 present the lowest values of failure per semester (X_3) and the highest values of academic performance values (X_5). This shows that these variables are the most influential in determining **WHO** will drop out. Furthermore, the Cox regression model to estimate the survival curves for the sample data and we need estimate **WHEN** a student is going to drop out. Given the sample data given in Table X we obtain that Student 1 and Student 3 have the best survival probabilities. Education experts indicate that the first two years (or four semesters) are the most critical and are where the highest dropout rates occur. Our proposal also demonstrates these facts. As we noted in Table XI, if the probability of surviving until the fourth semester is greater than 0.75, it is clear evidence that the student will not drop out. Then, combining the machine learning algorithms and survival analysis methods with the knowledge of education experts allows us to take steps over time to reduce university dropout rates.

V. DISCUSSION

After analyzing the results, RF is the classification algorithm with the best performance to identify **WHO** is at risk of dropping out. We obtained a good performance in all the classification models tested using the usual machine learning metrics and the Kappa coefficient. This shows that the Kappa value for RF is close to 0.6 that [34] define as a good value. Comparing to other models such as LR, SVM, NB, and DT, despite having similar performance according to AUC, their values with the Kappa coefficient were much lower.

TABLE XI
SUMMARY OF WHO AND WHEN PREDICTION OF STUDENTS DROP OUT.

Student	Student's dropout	$S(4\|X)$
1	0	0.884
2	1	0.012
3	0	0.775
4	1	0.593
5	1	0.518

To determine **WHEN** a student would drop out, we used survival analysis. First, we use non-parametric estimators as KM-estimator and NA-estimator. Based on these methodologies, it was shown by three-point of view (graphically, analytical, and by contrast hypothesis) that the variable associated with the student's gender is not relevant to predict the risk of dropout. On the other hand, the Cox model was applied to measure the effectiveness of the covariates on the level of risk. Besides, the Cox model shows that the variables X_3 (failures per semester) and X_4 (courses per semester) are more significant. X_3 represents the number of failures per semester and has the highest hazard radio value ($\text{HR}_3 = 86.26$). The effect of this variable is very predominant, as illustrated in Fig. 4. In contrast with other studies, the variable X_5 (GPA - academic performance) does not present a significant effect. X_2 (absences per course) has a hazard radio very close to one and consequently does not have a relevant influence on student dropout risk. Finally, in the function of the C-index, we obtain a good average performance to predict **WHEN** a student could drop out.

VI. LIMITATIONS AND FUTURE WORKS

We carried out this work with academic and demographic data. Although the results were good, we believe that it is possible to predict better who would drop out and when this would occur with more diverse data. The dropout of students with good academic performance may be due to other factors, such as having selected the wrong career. Economic factors could also be relevant to our analysis. For this reason, future works will seek to incorporate a greater diversity of data. These data must include socioeconomic, university, and psychology information that leads to more robust models. Moreover, it intends to compare different levels: for the whole university, other careers, and a student. In this way, it could be possible to apply similar actions for the university or groups. We will also seek to implement new methodologies that allow us to interpret when a student is going to drop out and implement explanation techniques to understand what if modifying features a student stops being a possible dropout.

VII. Conclusion

In this work, we presented a methodology to determine WHO would drop out and WHEN the drop would be. To answer these questions, we compared two approaches. The first uses traditional machine learning algorithms to predict WHO will drop out, and the second uses survival models to determine WHEN the withdrawal will occur. For both cases, we decided to create new variables guaranteeing the consistency of the values by adding time information. Although these approaches are generally approached separately, in this work, we find certain coincidences. For instance, C-index is a generalization of AUC, which is widely used to measure the performance of a machine learning classification algorithm. Another special attention is applying the Kappa coefficient to measure performance indices in machine learning classification models (**WHO**). Despite being little used in the literature, this index showed in our study that the RF had the best average performance. Based on the survival analysis (**WHEN**), it was shown that the number of failures per course is the most influential variable in the level of risk of dropping out. This variable has a very high Hazard ratio, given by $\mathrm{HR}_3 = 86.26$. In contrast, the GPA does not present a relevant effect on the risk of dropping out, $\mathrm{HR}_5 = 1.02$. Finally, our work reinforces the integration of probabilistic modeling approaches and machine learning algorithms.

Acknowledgment

This research was supported by the National Fund for Scientific and Technological Development and Innovation (Fondecyt-Perú) within the framework of the "Project of Improvement and Expansion of the Services of the National System of Science, Technology and Technological Innovation" [Grant #028-2019-FONDECYT-BM-INC.INV].

References

[1] Devesh Kapur and Megan Crowley. Beyond the abcs: Higher education and developing countries. *Center for Global Development Working Paper 139*, 2008.

[2] Vincent Tinto. Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research*, 45(1):89–125, 1975.

[3] Marek J. Druzdzel and Clark Glymour. What do college ranking data tell us about student retention? In *Intelligent Information Systems IV Proceedings of the Workshop held in Augustów*, pages 1–10, June 1995.

[4] Walber A. Ramos Beltrame and Oldair Luiz Gonçalves. Socioeconomic data mining and student dropout: Analyzing a higher education course in brazil. *International Journal for Innovation Education and Research*, 8(8):505–518, Aug. 2020.

[5] Caterina Balenzano, Giuseppe Moro, and Rosalinda Cassibba. Education and social inclusion: An evaluation of a dropout prevention intervention. *Research on Social Work Practice*, 29(1):69–81, 2019.

[6] Sattar Ameri, Mahtab J. Fard, Ratna B. Chinnam, and Chandan K. Reddy. Survival analysis based framework for early prediction of student dropouts. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, page 903–912, New York, NY, USA, 2016. Association for Computing Machinery.

[7] Artur Mesquita Barbosa, Emanuele Santos, and João P. P. Gomes. A machine learning approach to identify and prioritize college students at risk of dropping out. In *XXVIII Simpósio Brasileiro de Informática na Educação SBIE (Brazilian Symposium on Computers in Education)*, pages 1497–1506, Recife, Nov 2017.

[8] M. Solis, T. Moreira, R. Gonzalez, T. Fernandez, and M. Hernandez. Perspectives to predict dropout in university students with machine learning. In *2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI)*, pages 1–6, 2018.

[9] E. C. Medina, C. B. Chunga, J. Armas-Aguirre, and E. E. Grandón. Predictive model to reduce the dropout rate of university students in perú: Bayesian networks vs. decision trees. In *2020 15th Iberian Conference on Information Systems and Technologies (CISTI)*, pages 1–7, 2020.

[10] Juan C. Juajibioy. Study of university dropout reason based on survival model. *Open Journal of Statistics*, 6(5):908–916, 2016.

[11] Russell W. Rumberger and Scott L. Thomas. The distribution of dropout and turnover rates among urban and suburban high schools. *Sociology of Education*, 73(1):39–67, 2000.

[12] Deborah Pedraza and Mario Beruvides. The relationship between course assignments and academic performance: An analysis of predictive characteristics of student performance texas tech university. In *ASEE's 123rd Annual Conference & Exposition, New Orleans, LA*, June 2016.

[13] Ramon Henson. Expectancy beliefs, ability, and personality in predicting academic performance. *The Journal of Educational Research*, 70(1):41–44, 1976.

[14] Gilbert Greefrath and Wolfram Koepf. Is there a link between preparatory course attendance and academic success? a case study of degree programmes in electrical engineering and computer science. *International Journal of Research in Undergraduate Mathematics Education*, 3, 2017.

[15] Michalis Xenos, Christos Pierrakeas, and Panagiotis Pintelas. A survey on student dropout rates and dropout causes concerning the students in the course of informatics of the hellenic open university. *Computers & Education*, 39(4):361–377, 2002.

[16] Sergi Rovira, Eloi Puertas, and Laura Igual. Data-driven system to predict academic grades and dropout. *PLOS ONE*, 12(2):1–21, 02 2017.

[17] Stuart Palmer. Modelling engineering student academic performance using academic analytics. *International journal of engineering education*, 29(1):132–138, 2013.

[18] Ruba Alkhasawneh and Rosalyn Hargraves. Developing a hybrid model to predict student first year retention

in stem disciplines using machine learning techniques. *Journal of STEM Education*, 15(3):1557–5284, 2014.

[19] Umesh Pandey and Saurabh Pal. Data mining : A prediction of performer or underperformer usingclassification. *Journal of Computer Science and Technology*, 4:686–690, 04 2011.

[20] Ying Zhang, Samia Oussena, Tony Clark, and Hyeonsook Kim. Use data mining to improve student retention in higher education - a case study. In *Proceedings of the 12th International Conference on Enterprise Information Systems*, pages 190–197, 01 2010.

[21] Tuomas Tanner and Hannu Toivonen. Predicting and preventing student failure - using the k-nearest neighbour method to predict student performance in an online course environment. *International Journal of Learning Technology*, 5(4):356–377, 2010.

[22] Hashmia Hamsa, Simi Indiradevi, and Jubilant J. Kizhakkethottam. Student academic performance prediction model using decision tree and fuzzy genetic algorithm. *Procedia Technology*, 25:326 – 332, 2016.

[23] V. Hegde and P. P. Prageeth. Higher education student dropout prediction and analysis through educational data mining. In *2018 2nd International Conference on Inventive Systems and Control (ICISC)*, pages 694–699, 2018.

[24] Arto Vihavainen, Matti Luukkainen, and Jaakko Kurhila. Using students' programming behavior to predict success in an introductory mathematics course. In S. K. D'Mello, R. A. Calvo, and A. Olney, editors, *Proceedings of the 6th International Conference on Educational Data Mining*, pages 300–303, International, July 2013. International Educational Data Mining Society. Volume: Proceeding volume: ; The 6th International Conference on Educational Data Mining ; Conference date: 06-07-2013 Through 09-07-2013.

[25] Himabindu Lakkaraju, Everaldo Aguiar, Carl Shan, David Miller, Nasir Bhanpuri, Rayid Ghani, and Kecia L. Addison. A machine learning framework to identify students at risk of adverse academic outcomes. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 1909–1918, New York, NY, USA, 2015. Association for Computing Machinery.

[26] Haozhang Deng, Xuemeng Wang, Zhiyi Guo, Ashley Decker, Xiaojing Duan, Chaoli Wang, G. Alex Ambrose, and Kevin Abbott. Performancevis: Visual analytics of student performance data from an introductory chemistry course. *Visual Informatics*, 3(4):166–176, 2019.

[27] Bardh Prenkaj, Paola Velardi, Giovanni Stilo, Damiano Distante, and Stefano Faralli. A survey of machine learning approaches for student dropout prediction in online courses. *ACM Comput. Surv.*, 53(3), May 2020.

[28] Li Zhang and Huzefa Rangwala. Early identification of at-risk students using iterative logistic regression. In Carolyn Penstein Rosé, Roberto Martínez-Maldonado, H. Ulrich Hoppe, Rose Luckin, Manolis Mavrikis, Kaska Porayska-Pomsta, Bruce McLaren, and Benedict du Boulay, editors, *Artificial Intelligence in Education*, pages 613–626, Cham, 2018. Springer International Publishing.

[29] S. Massa and P. P. Puliafito. An application of data mining to the problem of the university students' dropout using markov chains. In Jan M. Żytkow and Jan Rauch, editors, *Principles of Data Mining and Knowledge Discovery*, pages 51–60, Berlin, Heidelberg, 1999. Springer Berlin Heidelberg.

[30] Jairo Coronado-Hernandez, Amelec Viloria, Mercedes Gaitán-Angulo, Nohora Mercado-Caruso, and Jose Arias-Pereze. Analysis of probability of dropout, continuation and graduation through markovian chains of university students in bolivar, colombia. *International Journal of Control Theory and Applications*, 9(44):257–263, 2016.

[31] Gonzalez-Campos, José Alejandro, Cristian Manuel Carvajal-Muquillaza, and Juan Elias Aspeé-Chacón. Modeling of university dropout using Markov chains. *Uniciencia*, 34:129 – 146, 06 2020.

[32] Schmitt Jeovani, Fini Maria Inês, Bailer Cyntia, Fritsch Rosangela, and Andrade Dalton Francisco de. Wwh-dropout scale: when, why and how to measure propensity to drop out of undergraduate courses. *Journal of Applied Research in Higher Education*, 32(2):429–454, 2020.

[33] Fernando Martínez-Plumed, Ricardo B.C. Prudêncio, Adolfo Martínez-Usó, and José Hernández-Orallo. Item response theory in ai: Analysing machine learning classifiers at the instance level. *Artificial Intelligence*, 271:18–42, 2019.

[34] Brett Lantz. *Machine learning with R*. Packt publishing ltd, 2013.

[35] David G. Kleinbaum and Mitchel Klein. *Survival Analysis*. Springer, Third edition, 2012.

[36] Frank E. Harrell, Robert M. Califf, David B. Pryor, Kerry L. Lee, and Robert A. Rosati. Evaluating the Yield of Medical Tests. *JAMA*, 247(18):2543–2546, 05 1982.