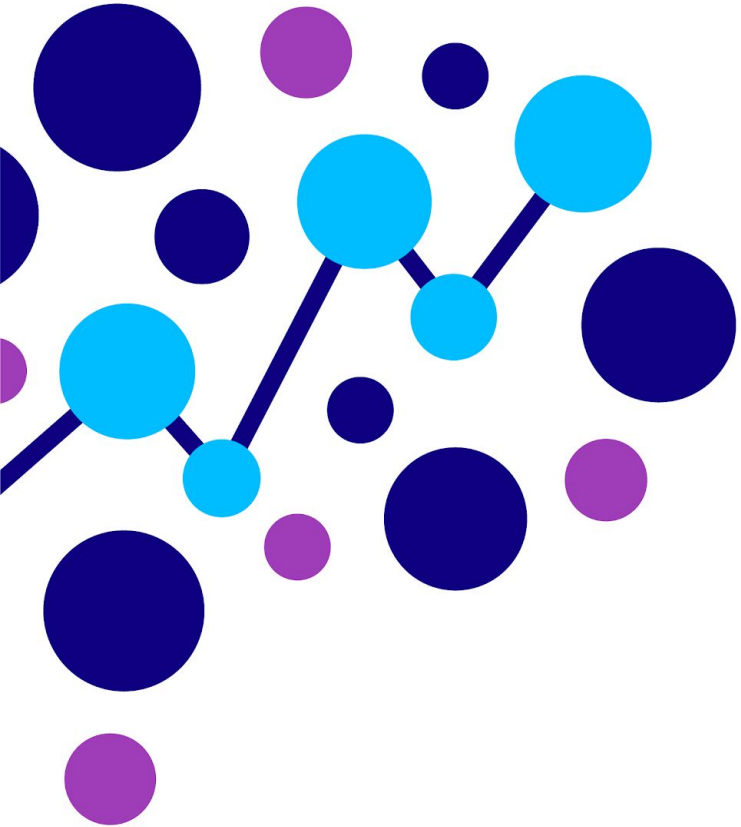


Models of Survival Analysis employed in the Prediction of University Dropout

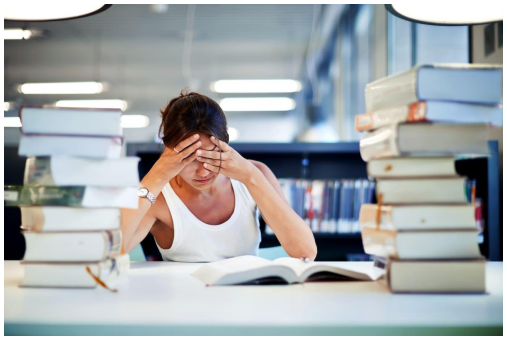
Daniel Alexis Gutierrez Pachas, Ph.D.
dgutierrezp@ucsp.edu.pe

Introduction

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.

Introduction

According to many works, **academic performance** is the main cause to be considered for the student **dropout**. However, this variable is not decisive in identifying students at risk of dropping out. A profound analysis must be able to deal with **multiple factors** and **temporal information**.

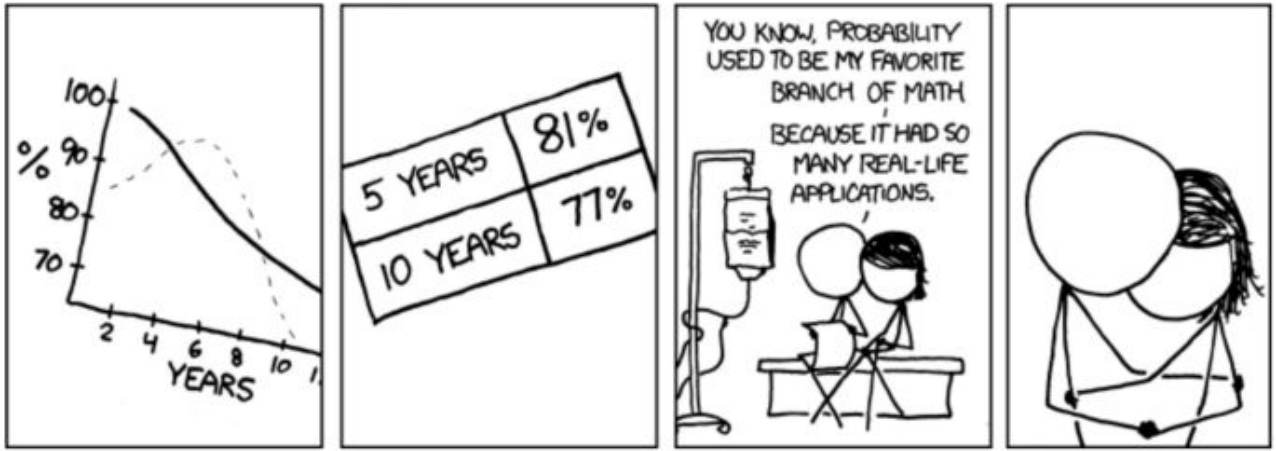


SDP problem

One of the first attempts to understand student dropout was Tinto's theoretical Dropout model. Subsequently, statistical tools were used to have a better understanding this educational problem.

Currently, machine learning computational techniques have been crucial to solve the **SDP problem** in various educational contexts. Even these techniques address other educational problems such as retention and academic performance of students.

Survival Analysis

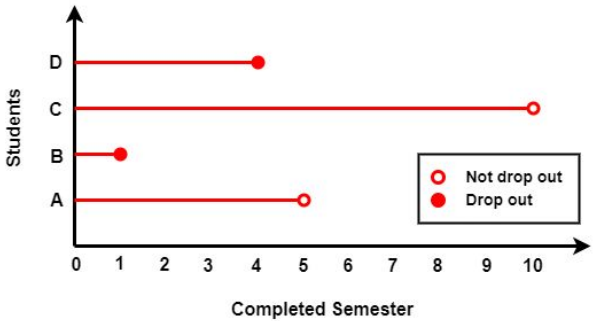


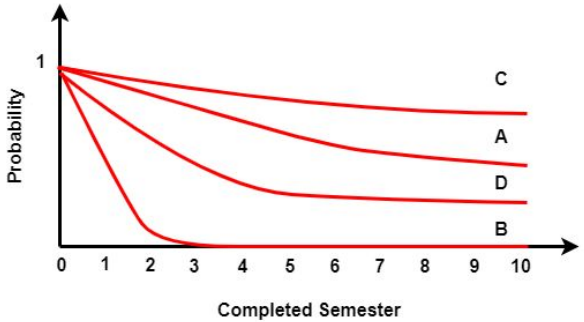
Survival Analysis

Survival analysis is used to analyze or predict **when an event** is likely to happen. It originated from medical research, but its use has greatly expanded to many different fields. For instance:

- Banks, lenders and other financial institutions use it to predict when a borrower will default.
- Engineers/manufacturers apply it to predict when a machine will break.
- Educational institutions use it to predict when a student will drop out of school.

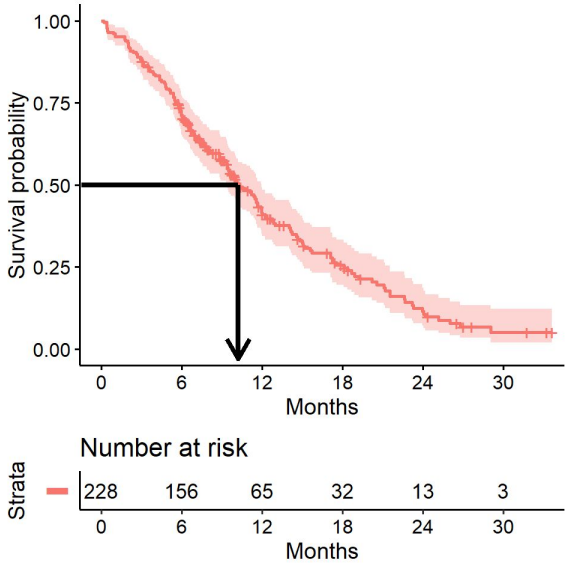
Survival Analysis





Survival Analysis

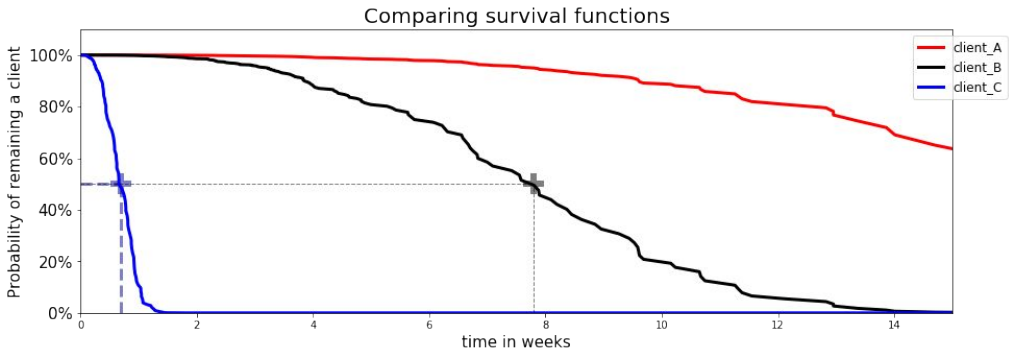
Survival function S(t) is the probability that the event has not occurred by some time t. This is **S(t)=1-F(t)=Prob(T≥t).**



Survival Analysis

The survival function given the feature vector x, is

$$S(t, x) = Prob(T \geq t \,|\, X = x)$$



We subdivide the time axis in J parts and calculate the risk score of a sample x by

$r(x) = \sum_{j=1}^{J} H(t_j, x)$. Therefore $r(x_A) < r(x_B) < r(x_C)$

Survival Analysis

Hazard function h(t) expresses the conditional probability that the event will occur within [t,t+dt[, given that it has not occurred before.

$$h(t) = \lim_{dt \to 0} \frac{Prob(t \leq T < t + dt \mid T \geq t)}{dt} = -\frac{d}{dt} log S(t)$$

CPH is a semi-parametric model define by

$$h(t|X) = h_0(t) e^{\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5}$$

$h_o(t)$ is the baseline hazard function. These methods are different from typical regression/classification because it depends on T.

Survival Analysis

In particular, non-parametric models do not use information from the variables but are used to have a global/partial knowledge of the data. Also, parametric models are based a probability distribution. In contrast, semi-parametric models is a combination between non-parametric and parametric models.

Survival Analysis

Harrell's C-index (also known as the **concordance index**) is a goodness of fit measure for models which produce risk scores. It is commonly used to evaluate risk models in survival analysis, where data may be censored.

$$\text{C-index} = \frac{\#\text{concordant pairs}}{\#\text{concordant pairs} + \#\text{discordant pairs}}$$

Survival Analysis

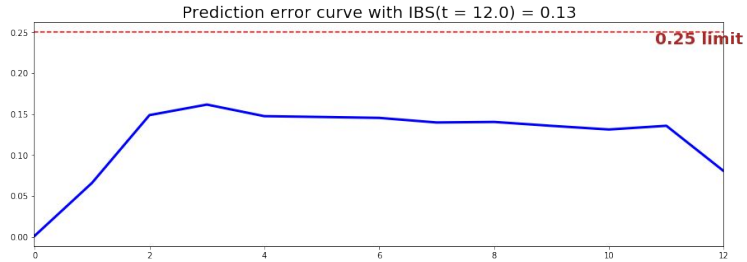
Harrell's C-index (also known as the **concordance index**) is a goodness of fit measure for models which produce risk scores. It is commonly used to evaluate risk models in survival analysis, where data may be censored.

$$\text{C-index} = \frac{\#\text{concordant pairs}}{\#\text{concordant pairs} + \#\text{discordant pairs}}$$

C-index is a generalization of the area under the ROC curve (AUC), commonly used in Classification Machine Learning methods. Similarly to the AUC, C-index=1 corresponds to the best model prediction, and C-index=0.5 represents a random prediction.

Survival Analysis

The **Brier score** evaluates the accuracy of a predicted survival function at a given time t; it represents the *average squared distances between the observed survival status and the predicted survival probability* and is always a number between 0 and 1, with 0 being the best possible value. In terms of benchmarks, a useful model will have a Brier score below 0.25. To assess the overall error measure across multiple time points, the Integrated Brier Score (IBS) is usually computed as well.



SDP problem as survival analysis model

Survival analysis methods could predict the dropout outcome an output **monotonically decreasing survival probabilities** to achiev consistent predictions along time.

$(X_{train}, T_{train}, E_{train})$

$(X_{test}, T_{test}, E_{test})$

Compare

(T_{test}, E_{test}) (T_{pred}, E_{pred}) = $f(X_{test})$

Methodology

- Data collection.
- Feature Engineering.
- Generation of predictive models.
- Visualization.
- Validation and Simulation.

Data Collection

Label name	Description	Type
MaskPerson_ID	Masked person identifier	Nominal
MaskStudent_ID	Masked student identifier	Nominal
NumStudent_ID	Number of student identities	Discrete numeric
Initial_CD	Initial enrolled curricular design	Nominal (IE1 or IE2 or CS1 or CS2)
Final_CD	Final enrolled curricular design	Nominal (IE1 or IE2 or CS1 or CS2)
Number_CD	Number of enrolled curricular designs	Discrete numeric
Changed_CD	Whether the student changed of curricular design or not	Nominal (Yes or No)
Admission_Sem	Admission semester to the university	Nominal (from 1999-01 to 2020-02)
Gender	Gender of student	Nominal (Male or Female)
Socioeconomic_Prov	Provenance location' socioeconomic level	Nominal (High or Medium or Low)
Final_GPA	Final Grade Point Average	Continuous numeric
Courses_Mean	Mean of enrolled courses per semester	Continuous numeric
Faults_Courses	Proportion of faulted courses in relation to the total number of enrolled courses	Continuous numeric
Absences_Courses	Proportion of faulted courses by the absence in relation to the total number of enrolled courses	Continuous numeric
Reservations_Courses	Proportion of reservations in relation to the total number of enrolled courses	Continuous numeric
NonRegular_Prop	Proportion of non-regular semesters in relation to the total number of enrolled semesters	Continuous numeric
Scholarship	Whether the student has a scholarship or not	Nominal (Yes or No)
Completed_Sem	Number of completed semesters	Discrete numeric (from 1 to 10)
Enrolled_Sem	Number of enrolled semesters	Discrete numeric
Student_Status	Student's current status	Nominal (Graduated or Regular or Reserved or Separated or Retired or Transferred)
Dropout	Dropout status	Nominal (Yes or No)

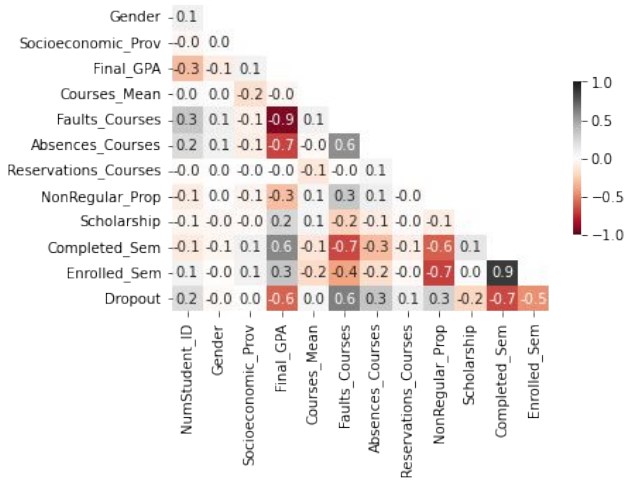
We obtained these data with the collaboration of the IT department, which was responsible for **masking sensitive data**. Thus, we do not have access to the student's name or personal information, preserving the student's identity.

Feature Engineering

Label name	Description	Type
MaskPerson_ID	Masked person identifier	Nominal
MaskStudent_ID	Masked student identifier	Nominal
NumStudent_ID	Number of student identities	Discrete numeric
Initial_CD	Initial enrolled curricular design	Nominal (IE1 or IE2 or CS1 or CS2)
Final_CD	Final enrolled curricular design	Nominal (IE1 or IE2 or CS1 or CS2)
Number_CD	Number of enrolled curricular designs	Discrete numeric
Changed_CD	Whether the student changed of curricular design or not	Nominal (Yes or No)
Admission_Sem	Admission semester to the university	Nominal (from 1999-01 to 2020-02)
Gender	Gender of student	Nominal (Male or Female)
Socioeconomic_Prov	Provenance location' socioeconomic level	Nominal (High or Medium or Low)
Final_GPA	Final Grade Point Average	Continuous numeric
Courses_Mean	Mean of enrolled courses per semester	Continuous numeric
Faults_Courses	Proportion of faulted courses in relation to the total number of enrolled courses	Continuous numeric
Absences_Courses	Proportion of faulted courses by the absence in relation to the total number of enrolled courses	Continuous numeric
Reservations_Courses	Proportion of reservations in relation to the total number of enrolled courses	Continuous numeric
NonRegular_Prop	Proportion of non-regular semesters in relation to the total number of enrolled semesters	Continuous numeric
Scholarship	Whether the student has a scholarship or not	Nominal (Yes or No)
Completed_Sem	Number of completed semesters	Discrete numeric (from 1 to 10)
Enrolled_Sem	Number of enrolled semesters	Discrete numeric
Student_Status	Student's current status	Nominal (Graduated or Regular or Reserved or Separated or Retired or Transferred)
Dropout	Dropout status	Nominal (Yes or No)

We use the number of enrolled semesters, the number of hours of absences, the number of approved courses, and the total number of courses to determine **proportional variables** between them, such for example Faults_Courses represent the proportion of faulted courses in relation to the total number of enrolled courses.

Visualization



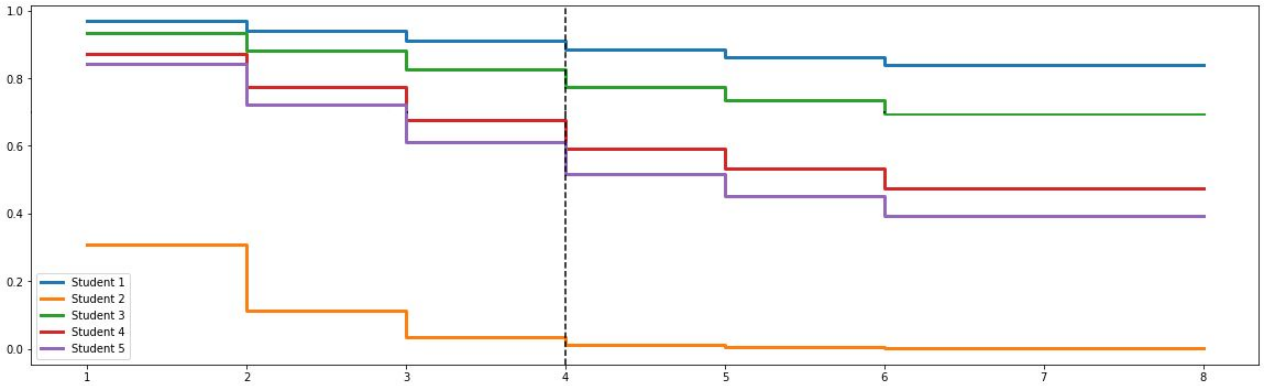
Gutierrez-Pachas, D.A.; Garcia-Zanabria, G.; Cuadros-Vargas, A.J.; Camara-Chavez, G.; Poco, J.; Gomez-Nieto, E. How Do Curricular Design Changes Impact Computer Science Programs?: A Case Study at San Pablo Catholic University in Peru. *Educ. Sci.* **2022**, *12*, 242. https://doi.org/10.3390/educsci12040242

Generation of predictive models

We define a dropout student as a student who is not a graduate and does not record activity after a specific time. Dropout' status is the "event variable" (E=Dropout). Also, "temporal variable" corresponds to a student's time active, (T=Completed_Sem). Also, X denotes the vector of "predictor variables," and Y the "target variable"; however, **Y has a different definition for each technique**

- Y= Dropout for Classification models.
- Y= (Completed_Sem,Dropout) for Survival models.

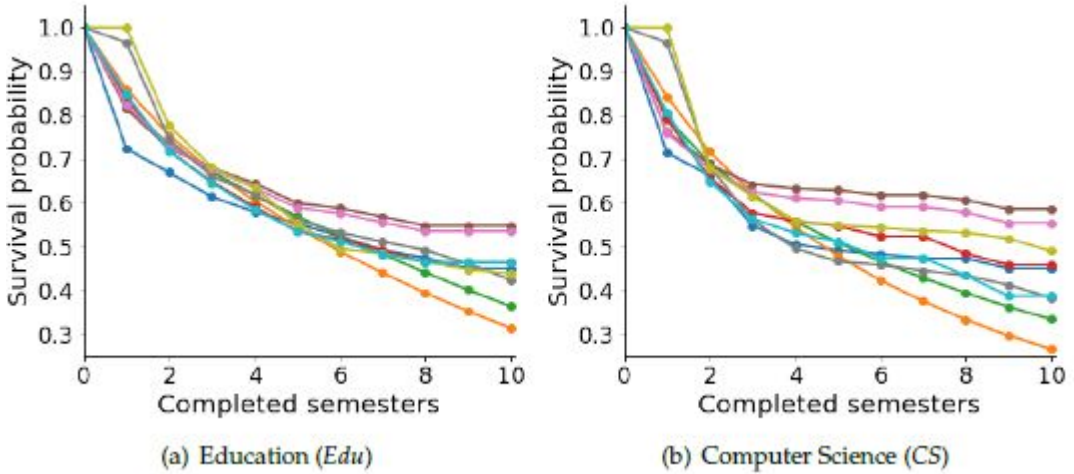
Validation and Simulation



According this value, if $r(x_A)<r(x_B)$, it means that the student with features x_A has less risk of dropping out than the student with features x_B

Gutierrez-Pachas, D.A.; Garcia-Zanabria, G.; Cuadros-Vargas, A.J.; Camara-Chavez, G.; Poco, J.; Gomez-Nieto, E. A comparative study of WHO and WHEN prediction approaches for early identification of university students at dropout risk. Latin American Computing Conference, *2021.*

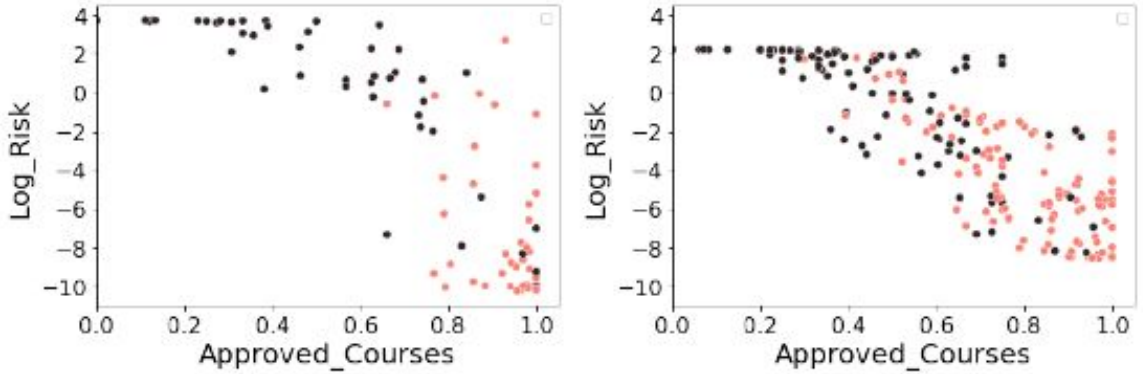
Validation and Simulation



(a) Education (*Edu*) (b) Computer Science (*CS*)

Figure 2. Comparison of predicted survival curves. The Actual curve is displayed in blue (—), while the predicting methods are: Weibull (—), Gompertz (—), CPH (—), RSF (—), CSF (—), MTLR (—), N-MTLR (—), and DeepSurv (—).

Validation and Simulation



(a) Education (*Edu*) (b) Computer Science (*CS*)

Figure 3. Scatter plot between the proportion of approved courses and the logarithm of risk score. We highlighted a student dropout in black (●), otherwise in pink (●).

Proyecto
Concytec
Banco Mundial

Universidad Católica San Pablo
Arequipa - Perú

ProCIENCIA

CONCYTEC
CONSEJO NACIONAL DE CIENCIA, TECNOLOGÍA E INNOVACIÓN TECNOLÓGICA

BANCO MUNDIAL
BIRF • AIF | GRUPO BANCO MUNDIAL

Universidad Católica San Pablo 25 AÑOS

THANK YOU

Daniel Alexis Gutierrez Pachas, PhD
dgutierrezp@ucsp.edu.pe