

UNIVERSIDAD ADOLFO IBÁÑEZ

TESIS DE MAGÍSTER

---

**Analysis of first-year university student dropout through  
machine learning models: A comparison between  
Universities**

---

*Autor:*  
Diego Opazo Inarejo

*Profesor guía:*  
Sebastián Moreno

*Comisión:*  
Jordi Pereira  
Eduardo Álvarez

*Tesis para optar al título de:*  
*Ingeniero Civil Informático*  
*Ingeniero Civil Industrial, mención Business Analytics*  
*Grado de Magíster en Ciencias de la Ingeniería, mención Tecnologías de la Información*  
*de la*

**Facultad de Ingeniería y Ciencias**

July 19, 2021

**UAI**

FACULTAD DE  
INGENIERÍA Y  
CIENCIAS

UNIVERSIDAD ADOLFO IBÁÑEZ

## *Abstract*

Facultad de Ingeniería y Ciencias

Tesis para optar al título de:  
Ingeniero Civil Informático  
Ingeniero Civil Industrial, mención Business Analytics  
Grado de Magíster en Ciencias de la Ingeniería, mención Tecnologías de la Información

### **Analysis of first-year university student dropout through machine learning models: A comparison between Universities**

by Diego Opazo Inarejo

The student dropout, defined as the abandonment of a high education program before obtaining the degree without reincorporation, is a problem that affects every higher education institution. In this work, we used machine learning models over two Chilean universities to predict first-year engineering student dropout and analyze the variables that could affect the students. The results showed that instead of joining the datasets into a single dataset, it is better to apply a model per university. Also, the best model per dataset was gradient boosting decision trees. Finally, the interpretative models determined that a higher score in almost any test decreases the probability of dropout, being the most significant variable the mathematical test. One exception is the language test, where a higher score increases the probability of dropout.

**Palabras clave:** Machine Learning, Classification, first-year student dropout, universities.

# Contents

|                                                     |           |
|-----------------------------------------------------|-----------|
| <b>Abstract</b>                                     | <b>i</b>  |
| <b>1 Problem Description</b>                        | <b>1</b>  |
| <b>2 State of the art</b>                           | <b>3</b>  |
| 2.1 Data mining models applied to dropout . . . . . | 4         |
| 2.1.1 Decision trees . . . . .                      | 4         |
| 2.1.2 Logistic regression . . . . .                 | 5         |
| 2.1.3 Naive Bayes . . . . .                         | 5         |
| 2.1.4 K-Nearest Neighbors (KNN) . . . . .           | 5         |
| 2.1.5 Neuronal Networks . . . . .                   | 5         |
| 2.1.6 Comparing models . . . . .                    | 6         |
| <b>3 Hypothesis and objectives</b>                  | <b>7</b>  |
| 3.1 Hypothesis . . . . .                            | 7         |
| 3.2 Objectives . . . . .                            | 7         |
| <b>4 Methodology</b>                                | <b>8</b>  |
| <b>5 Exploratory Data Analysis</b>                  | <b>11</b> |
| 5.1 Universidad Adolfo Ibáñez . . . . .             | 11        |
| 5.2 Universidad de Talca . . . . .                  | 13        |
| 5.3 Unification of both datasets . . . . .          | 16        |
| <b>6 Results and discussion</b>                     | <b>17</b> |
| 6.1 Both-universities-dataset results . . . . .     | 17        |
| 6.2 Universidad Adolfo Ibáñez results . . . . .     | 20        |
| 6.3 Universidad de Talca results . . . . .          | 24        |
| 6.4 Universidad de Talca all variables . . . . .    | 28        |
| 6.5 Discussion . . . . .                            | 30        |
| <b>7 Conclusion</b>                                 | <b>35</b> |
| <b>A Machine learning models</b>                    | <b>40</b> |
| A.1 K-Nearest Neighbors . . . . .                   | 40        |
| A.2 Support vector machine . . . . .                | 40        |
| A.3 Decision Trees . . . . .                        | 40        |
| A.4 Random forest . . . . .                         | 41        |
| A.5 Gradient boosting decision tree . . . . .       | 41        |
| A.6 Naive Bayes . . . . .                           | 41        |
| A.7 Logistic Regression . . . . .                   | 42        |
| A.8 Artificial Neuronal Networks . . . . .          | 42        |

|          |                                              |           |
|----------|----------------------------------------------|-----------|
| <b>B</b> | <b>Variable description</b>                  | <b>43</b> |
| B.1      | Universidad Adolfo Ibáñez . . . . .          | 43        |
| B.2      | Universidad de Talca . . . . .               | 50        |
| <b>C</b> | <b>Selection of parameters and variables</b> | <b>60</b> |
| C.1      | K-Nearest Neighbors (KNN) . . . . .          | 60        |
| C.2      | Support vector machine (SVM) . . . . .       | 61        |
| C.3      | Decision tree . . . . .                      | 63        |
| C.4      | Random forest . . . . .                      | 65        |
| C.5      | Gradient boosting decision tree . . . . .    | 67        |
| C.6      | Naive Bayes (NB) . . . . .                   | 70        |
| C.7      | Logistic regression . . . . .                | 73        |
| C.8      | Neural network . . . . .                     | 75        |



## Chapter 1

# Problem Description

Education is one of the most important factors in the development of a country. A better and extensive education helps to achieve better levels of social wellness and economic growth. At the same time, education decreases social inequalities and promotes social mobility. Finally, it also promotes science, technology, and innovation. In summary, it helps to build a better society. According to a report from UNESCO in 2015<sup>1</sup>, the global number of students in high education has grown from 28.5 million in 1970, to 196 million in 2012. However, not all these students have necessarily finished their degree, as an important part of them drop out.

The student dropout, defined as the abandonment of a high education program before obtaining the degree without reincorporation ([1]), is a problem that affects every higher education institution. In the United States, the overall dropout rate for undergraduate college students is 40%<sup>2</sup>. In the European Union, the following countries have the lowest dropout rates: United Kingdom, Norway, and France<sup>3</sup> (16%, 17%, and 24% respectively), while Italy has the highest dropout rate (33%), followed by the Netherlands (31%)<sup>4</sup>. In Latin America, on average, 50% of the people between the ages of 25 and 29 who started a university degree did not complete their studies<sup>5</sup>.

There are different types of dropouts, and each of them can be analyzed in different ways. Even though a student can drop out of college at any time during his career, most dropouts happen during the first year. In the United States, until 2018, approximately 30% of college freshmen drop out before their sophomore year<sup>6</sup>. In the United Kingdom, 6.3% of young full-time students dropped out during the first year in 2016-17<sup>7</sup>. In Latin America, Colombia had a 36% of student dropout in the first year in 2013<sup>8</sup>, while Chile had a 21% in 2019<sup>9</sup>. For example, Universidad Adolfo Ibáñez and Universidad de Talca, the universities analyzed for this thesis, have a 12% and 15% of first-year student dropout<sup>10</sup>.

Studies have determined multiple factors for first-year dropouts ([2, 3, 4, 5]). According to [3] and [2], the first year of high education is a critical year to determine if a student will drop out or not. However, other studies, like [4], determined that students drop out due to vocational problems; while [5] identified that school performance is one of the best predictors for dropout. As can be observed, the first-year dropout is a complex and important problem to be solved. Moreover, it is important to consider that conclusions from other countries are not necessarily applicable to the rest of the world,

---

<sup>1</sup><https://unesdoc.unesco.org/ark:/48223/pf0000234743>

<sup>2</sup><https://educationdata.org/college-dropout-rates/>

<sup>3</sup><https://www.timeshighereducation.com/news/uk-has-lowest-drop-out-rate-in-europe/2012400.article>

<sup>4</sup><https://www.timeshighereducation.com/news/uk-has-lowest-drop-out-rate-in-europe/2012400.article>

<sup>5</sup><https://openknowledge.worldbank.org/handle/10986/26489>

<sup>6</sup><https://educationdata.org/college-dropout-rates/>

<sup>7</sup><https://www.studyinternational.com/news/dropping-out-university/>

<sup>8</sup><https://openknowledge.worldbank.org/handle/10986/26489>

<sup>9</sup><https://www.mifuturo.cl/informes-retencion-de-primer-ano/>

<sup>10</sup><https://www.mifuturo.cl/>

because of the different cultures. For example, in Chile, first-year students do not live inside the buildings of the campus as in the United States of America.

In this thesis, we will compare the learned patterns from machine learning models over two different universities to analyze their dissimilarities. Posterior analysis of these models will determine if the same dropout behavior pattern can be observed in both universities.

## Chapter 2

# State of the art

The problem of university student dropout has been studied for many years, approaching the problem from different perspectives. One of the first and most popular works have been [2] and [3] adaptation models. Both of them argue about how adaptation and social integration affect the decision of dropout. [2] based his model on Durkheim's theory of suicide, being the dropout the result of a complex social process that includes family and previous educational background, academic potential, normative congruence, friendship support, intellectual development, educational performance, social integration, satisfaction, and institutional commitment. [3] formulated a model considering factors like student adaptation, compromise with the institution, and previous academic performance. Similarly, [6] proposed the student attrition model, arguing that student dropout depends on factors that affect directly the student (factors external to the university), affecting their sense of wellness. On the contrary, [7] proposed a mixed model, where the main factors are related to the quality of the institution, the security in the choice of the career, or the existence of scholarships.

The previous models served as a base for new theoretical investigations. [4], for example, clusters the theoretical explanatory models into four different groups. **The adaptation model:** describing the lack of an individual's integration into the context. **The structural model:** the university structure, including political, economic, and social, that influences students to dropout. **The economic model:** student choice of an alternative way to invest time, energy, and resources that could offer greater benefits in the future. **The psycho-pedagogic model:** a mix of different factors from the adaptation and structural models, plus other dimensions of a psycho-educative nature, like learning strategies.

Other studies, based on mainly interviews and manual analysis, have reached diverse conclusions. [5] concluded that previous academic performance variables are the best predictors of the final performance, but leaving that aside, time management is also important. [8], from the implementation of a survey applied to young deserters, determined that dropouts are mostly related to socioeconomic or individual problems. Similarly, [9] inferred that the main cause of dropout is the lack of funding of the students. Proposing that the action of the welfare department should be more proactive to guide students in the financing options that they can access. On the contrary, [10] believes that the causes of dropout for one student may not be as relevant to another, meaning that each student dropouts for a different reason.

New studies applied mathematical models to predict student dropout, from genetic algorithms to multivariate and survival analyses. These added a new perspective, generating new conclusions. [11] collected data from a web-based system and used a genetic algorithm to predict student performance. Specifically, it analyzes the time students spent on resources from the university web system, improving the accuracy over previous classifiers by about 10%. [12] applied correlational analysis, relating the student dropout with four different variables: previous academic performance, first-year college performance, attendance, and date of enrollment (students whose date of insertion in the degree is after the start of the academic year have a higher dropout rate). [13] used a static econometric

model concluding that previous academic performance and funding are some of the best indicators for dropout. [14], using rule-based knowledge discovery, identified relevant causes for student dropout in the first and second year of a school of engineering. For first-year students, the most important aspects are: who funds students university studies and the number of years from the end of secondary school to university entrance. However, these factors change for second-year students, where a common dropout denominator is the “number of not subjects not attended as full-time students”, i.e., full-time students are most likely to finish their career. Recently, [15], in a world where the massification of technology has been mostly adopted by the young population, included the role of procrastination as a factor in student dropout. They show an association between high levels of procrastination and low academic performance in students. It also showed that procrastination can be evaluated with entry tests, and can be overcome by training.

Statistical models have also been applied for this type of analysis. [16] used maximum likelihood probit models to estimate the effects of specific factors that may lead to student dropout, concluding that better results on national high school exams substantially reduce the risk of dropout, and that female students also have a relatively lower estimated probability. [17] used the multivariate analysis technique to determine factors that affect university dropout, using a questionnaire to collect data. Concluding that the factors that most affect student dropout would be: economical (individual and family), institutional (management, institutional intervention, and monitoring student), mental (psychosocial and family support), and personal (motivational and social relationship).

Finally, more complex statistical models such as survival models have also been applied to analyze this problem. [18] analyzed socioeconomic characteristics and personal factors related to dropouts using duration analysis (the dropout is analyzed as a process in time, evaluating which students are most likely to abandon or graduate). Among her conclusions was that men, previous dropout students, and working students are more likely to drop out; besides the following variables: education of the parents and the type of secondary school (the family environment). [19], employed a discrete-time competing risks survival model to identify risk factors associated with high education dropout in the Pontificia Universidad Católica de Chile (PUC), and extended previous methods by handling covariate selection within a Bayesian variable selection framework. They concluded that there were strong levels of heterogeneity between different programs of the PUC; hence, building a common model for the entire university was not recommended.

## 2.1 Data mining models applied to dropout

The data mining process corresponds to the application of machine learning models on large amounts of historical data, finding new patterns. Lately, institutions have started to collect their data to generate value from it, which improves the quality of the predictive models. In this subsection, we analyze the application of typical machine learning models for dropout, to see a general definition of each model see [A](#).

### 2.1.1 Decision trees

The decision trees are structures used to classify based on decisions, where each leaf determines a class label ([20]). One of the firsts decision tree models applied to dropout corresponds to [21]. This work compared multiple training processes: ID3, C4.5, and ADT for dropout, concluding that the best decision tree is the ADT, with a precision rate of 82.8%, but without informative conclusions. Similarly, [22] applied different decision tree training algorithms to predict student dropout at Simón Bolívar University (Colombia). Even though it mentioned that decision trees are a good model, it did not arrive at any conclusion for features, as different training algorithms selected dissimilar variables. Finally, [23] determined that decision trees with parameter optimization result in a better

precision compared to other models. Moreover, determined three significant variables that could explain dropouts: grades, years of advancement in the career, and admission test university scores.

### 2.1.2 Logistic regression

One of the main advantages of logistic regression, first used by [24] as a probability model, is its interpretability, where each variable is associated with a parameter indicating its relevance. [25] provided a methodology to apply the logistic regression model to the student dropout problem, focusing on providing basic information to educational researchers. [26] analyzed dropout in Chilean higher education, concluding that the dropout is related to socioeconomic level, previous academic performance, score in the university admission test, academic scholarships, and financial credits. Surprisingly, one of the strongest correlations with persistence in high education is given by government financial credits and scholarships, implying the important financial restrictions in Chilean high education. Finally, [27] analyzed over seventeen variables determining seven significant variables: gender, time of study (day or evening), age group, school of origin, lives with family, score in the university admission test, and father's occupation; being admission test score the most important feature.

### 2.1.3 Naive Bayes

The Naive Bayes is a probabilistic model, based on the Bayes theorem, which also could be interpreted ([28]). [29] analyzed data from Dr. R. M. L. Awadh University, India, identifying highly correlated factors with the previous academic performance, living location, language of teaching (classes in a mix with mother tongue or English), mother's education, students habits, family annual income, and student's family status. Later, [30] applied this model (naive Bayes) over data from the Amrita School of Arts & Sciences, to predict early dropout. In this study, the most relevant variables were: academics, demographic, psychological, and health factors.

### 2.1.4 K-Nearest Neighbors (KNN)

KNN classifies an object based on the vote of its  $K$  closer neighbors, which are determined by a distance function ([31]). To date, there is not much research dedicated exclusively to predict university student dropout or similar problems. [32] used KNN to predict student performance in a touch-typing online course. Specifically, it identified at an early stage of the course those students who have a high risk of failing, using variables collected from course lessons, like the typing speed, accuracy, time spend in the lesson, and exam attempts. Recently, [33] used this model to predict school student dropout based on welfare-related variables, like parent's involvement, education, and annual income.

### 2.1.5 Neuronal Networks

Even though the great power of neural networks ([34]), they are considered black box; implying that the learned parameters can not be interpreted ([35]). [36] used student surveys, telephone interviews, and administrative data related to predict student dropout in a school of medicine. The final characteristics can be summarized in: personal, parents features, location, previous academic performance, university admission test scores, obtaining a precision between 65% and 84%. A posterior sensitive analysis determined that the most important variables are: family education, school origin, lack of pre-university guidance, study with friends, and motivation. Another example of this type of model is the work of [37], where a multilayer perceptron obtained a prediction rate of 96.3% (96.8% using a radial base function), using variables that can be summarized in: is or is not a parent, knowledge in software used in the university major, family commitment, adaptation to the university, university ranking and student's perspective on his or her integration into the labor market.

### 2.1.6 Comparing models

Besides previously described works, several investigations have used and compared more than one model to predict university dropout. [1] compared decision trees, neuronal networks, support vector machine, and logistic regression, obtaining the best performance with support vector machine. It also concluded that the most important predictors are: past and present educational success and financial help. [38] analyzed Engineering careers dropout from Universidad de Las Americas, comparing neuronal networks, decision trees, and K-median with the following variables: score in the university admission test, previous academic performance, age, and gender. Unfortunately, the research had no positive results, due to the unreliable data. [39] compared decision trees, Bayesian networks, and association rules, obtaining the best performance with decision trees, determining the following important variables: previous academic performance, origin, and age of student when he/she entered the university. Also, it identified that during the first year of the degree is where containment, support, tutoring, and all of the activities that improve the academic situation of the student upon entering the university acquire greater importance. Lately, two similar works ([40, 41]) used Bayesian networks, neuronal networks, and decision trees determining that the most influential variable were: university admission test scores and economical benefits (scholarships and credits). Finally, [42] compare logistic regression against decision trees, obtaining slightly better results with the decision trees than the logistic regression, concluding that the most relevant single factors to predict study success and dropout are combined features such as the count and the average of passed and failed examinations or average grades.

## Chapter 3

# Hypothesis and objectives

### 3.1 Hypothesis

- Patterns for first year dropout students differ among universities. A general model for dropout students, considering all universities, will learn different patterns than single models per university.

### 3.2 Objectives

- To apply the data mining process to the datasets from Universidad de Talca and Universidad Adolfo Ibáñez.
- To use the following machine learning models: KNN, SVM, decision tree, random forest, gradient boosting decision tree, naive Bayes, logistic regression, and neural network.
- To find the significant variables that differ between a general model (one dataset with both universities) and local models (one dataset per university).

## Chapter 4

# Methodology

In this thesis, we will compare the learned patterns from machine learning models over different universities to analyze their dissimilarities. We will create multiple models to forecast the drop out in engineering schools using datasets from two Chilean Universities (Universidad Adolfo Ibáñez and Universidad de Talca). Posterior analysis of these models will determine if the same dropout behavior pattern can be observed in both universities.

To achieve this goal, we separate this thesis into four sections. First, an exploratory data analysis, to understand the data and their variables. This analysis also includes data pre-processing where data points and variables are cleaned. Second, the implementation of machine learning models, including a tuning process of their hyperparameters. This section also includes the selection of the best variables for each model based on forward selection. Third, a data evaluation process, where all models are evaluated using k-fold cross-validation. Finally, an interpretation process, where the patterns or learned parameters are analyzed to generate new information applicable to future incoming processes.

In the exploratory data analysis, we gather initial information from the data through the description of each variable, showing their distributions, possible values, and missing data. In this process, we clean the data by discarding variables gathered after first-year, since they can not be used for first-year dropout prediction. Other unnecessary variables are also deleted, as well as problematic observations, like older observations with several missing variables. In some variables, we grouped possible values to improve quality and reduce complexity. We also analyze missing data points, searching for a possible pattern in their distribution to be replaced. In the case of outliers, most of them are eliminated indirectly when deleting older data. Finally, we analyze posterior distributions conditioned in the dropout variable, searching for relationships and initial explanatory variables.

All models were implemented using python and the libraries scikit-learn and Keras. In this implementation, we tune the hyperparameters and select the variables included in each model to improve the performance. The selection of variables is done using forward selection. Forward selection starts with an empty model, and, at each iteration, it selects among all variables, the one that obtains the best performance. This process is iterated until all variables are used or the results do not improve.

To evaluate the models, we will estimate the mean and standard deviation error through K-fold cross-validation on different measures (accuracy,  $F_1$  score for both classes, and AUC). K-fold cross-validation helps us to estimate the error distribution by splitting the datasets into  $k$  folds. Then,  $k - 1$  folds are selected for training and tested in the other fold. This process is repeated until all folds are used for testing, and the error estimation is given by the average over error folds. Besides, considering that we will model student dropout, there is likely to be a big difference in the proportion of data between them, leading to an unbalanced data problem, which will be minimized through undersampling. Specifically, the majority class is reduced through random sampling, so that the proportion between the majority and minority class is the same. To combine both methods (K-fold



cross-validation with an undersampling technique), we will apply the undersampling technique over each trainset produced after the K-fold split and then evaluate in the original test fold. With that, we avoid possible errors of double-counting duplicated points in the test sets when evaluating them.

We will measure the performance of each model using the following measures: accuracy,  $F_1$  score for both classes, and AUC. However, to fully understand the  $F_1$  score is also necessary to understand recall and precision. These measures are based on the confusion matrix, which is a table showing the performance of a classification model using four values: True Positives (TP) is the number of positive points classified correctly; True Negatives (TN) is the number of data points correctly classified as negative; False Positives (FP) is the number of data points incorrectly classified as positive; and False Negatives (FN) which is the number of data points incorrectly classified as negative.

The accuracy (equation 4.1) is one of the most basic measures, it indicates the percentage of correctly classified points, which is the number of TP plus TN, over the total number of data points. The accuracy varies between 0 and 1, where a high accuracy implies that the model can predict most of the data points correctly. However, this measure behaves improperly when a class is biased because high accuracy is achievable labeling all data points as the majority class.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (4.1)$$

To solve this problem we will use other measures that avoid the TN reducing the impact of biased datasets. The recall (equation 4.2) is the number of TP over the total points which belong to the positive class (TP+FN). The recall varies between 0 and 1, where a high recall implies that most of the points which belong to the positive class are correctly classified. However, we can have a high value of FP without decreasing the recall.

$$Recall = \frac{TP}{TP + FN} \quad (4.2)$$

The precision (equation 4.3) is the number of TP over the total points classified as positive class (TP + FP). The precision varies between 0 and 1, where a high precision implies that most of the points classified as positive class are correctly classified. With precision, it is possible to have a high value of FN without decreasing its value.

$$Precision = \frac{TP}{TP + FP} \quad (4.3)$$

To resolve the problems from recall and precision, we use the  $F_1$  score (equation 4.4). The  $F_1$  score is the harmonic average of the precision and recall, seeking a balance between them, improving the score on unbalanced data. It is calculated based on TP, FP, and FN. The  $F_1$  score varies between 0 and 1, where a high  $F_1$  score implies that the model can classify the positive class and (most importantly) generates a low number of FN and FP. Even though, TP is associated with the class with fewer labels, we will report the  $F_1$  score using both classes as TP, avoiding misinterpretation of the errors.

$$F_1score = \frac{2 * TP}{2 * TP + FP + FN} \quad (4.4)$$

In addition to the previous measures, we use AUC. The AUC (area under the curve) is obtained from the ROC curve, and measure the performance of the model at different classification thresholds. The ROC curve plots the true positive rate (4.5 left) vs the false positive rate (4.5 right) for all possible classification thresholds; visualizing the performance of the model. The AUC varies between 0 and 1,

where a high AUC implies better performance. Note that an AUC higher than 0.5 implies a prediction better than an aleatory decision.

$$TPR = \frac{TP}{TP + FN} \quad FPR = \frac{FP}{FP + TN} \quad (4.5)$$

Finally, we will analyze the patterns and learned parameters from some models. Decision tree, random forest, gradient boosting decision tree, logistic regression, and naive Bayes are interpretable models, meaning that we can extract pattern behaviors from each of these models. This is especially relevant for dropout students, where early action can be developed to avoid their dropout. For example, for a decision tree model, we will look at the feature importances, and at the attributes and conditions used in each node, where the path for each leaf can be translated into a set of rules indicating a common behavior of the students. Similar to the decision tree, for a random forest model, we will look at an average of the feature importances of the trees. In the case of a logistic regression model, we will look at the learned parameters ( $\beta_i$  values) to identify the most relevant attributes (higher absolute values are associated with the most relevant variables related to dropout). Moreover, an increment of one unit in a variable associated with  $\beta_i$  implies a proportional increment in the odds of the final probability. Finally, in the case of a naive Bayes model, we will analyze the conditional distributions given the dropout outcome of the student to identify the most relevant attributes and generate new behavior patterns.

## Chapter 5

# Exploratory Data Analysis

We acquired and analyzed the data from two Chilean universities, Universidad Adolfo Ibáñez and Universidad de Talca. In this section, we summarize the results from [B](#). For both datasets, we describe and analyze each variable, eliminating some of them and, in some cases, some specific data points. Finally, we also generate a single dataset including both universities.

### 5.1 Universidad Adolfo Ibáñez

Universidad Adolfo Ibáñez data had 31,714 observations and 40 variables. Each observation corresponds to a student from Universidad Adolfo Ibáñez. For this thesis, we considered only the engineering students, reducing the original dataset to 8,416 observations. However, this dataset contained a lot of null values and variables that did not contribute to the prediction of dropout. Then, we performed a data cleaning, deleting the unnecessary variables and part of the null values.

First, we discarded variables that do not contribute to the prediction of university first-year dropout. Five variables were eliminated before the analysis because their information was related to data quality, students that did not register, or do not influence dropout:

- **SIN NOTAS:** Students without information (data quality variable).
- **PROGRAMA:** Student program, since we considered only engineering students, the student program has a single value, engineering.
- **RETRACTOS ADMIN ESPECIAL:** Students who were interested in the university through a special admission, but did not sign-up.
- **TIPO COLEGIO 2:** School education type of the student. The variable distribution contains practically a single value.
- **FECHA NACIMIENTO:** Year of birth, it does not influence in the dropout decision (see [B](#)).

This process reduced the dataset to 8,416 observations and 35 variables. A second set of variables were eliminated because their information is gathered after the first year is completed, therefore not useful for first-year dropout prediction:

- **ESTADO ALUMNO AL TÉRMINO DE LOS 4 AÑOS DE PREGRADO:** Active student after 4 years.
- **FECHA TITULACIÓN:** Graduation date of the student.
- **cantidadReprobAño2:** Failed subjects in second-year.
- **promedioAño1:** First-year grades.
- **promedioAño2:** Second-year grades.

- **dobleTitul**: Student enrolled in two or more programs.
- **asigReprob**: Number of failed subjects in first-year.
- **INTERCAMBIO**: Student went one semester abroad.

The previous processes reduced variables, obtaining 8,416 observations and 27 variables. However, more extensive analysis and actions were still necessary to increase the data quality:

- The observations with COHORTE value equal to 2019 were eliminated, because we do not have the status of the student after the first year.
- Semestres, the variable that indicates the semester when the student drop out of school, was transformed to a 0/1 variable, being 1 when a student dropout in first-year, and 0 in other cases.
- The variable COMUNA, which indicates the commune where the student comes from, was reduced from 148 values to just 16 values: the 15 most common values, and others, which gather all other communes.
- The variable REGIÓN, which indicates the region where the student comes from, was reduced from its 16 possible values to 4. The reduction considered the 2 most frequent values (Región Metropolitana and Región de Valparaíso), in addition to the values North and South, containing the rest of the regions (regions at the north of Valparaíso and south of Región Metropolitana).
- The variables PSU CIENCIAS and PSU HISTORIA (test scores) were combined in the new PSU OPCIONAL variable, containing the maximum value between them (if both exist). This process was applied as only one test is required to report for admission.
- The variable ID ALUMNO was not eliminated for future checking.

These actions reduced the observations to 5,733 and 26 variables. Unfortunately, multiples variables still have some null values in the dataset. A first analysis showed that most of the null values are observations previous to 2012 COHORTE (10 of the 26 remaining variables, including student scores). The elimination of these observations reduced the data set to 3,750 students. Finally, we eliminated some variables with a high percentage of null values, including: MÓDULO PRUEBA CIENCIAS, NIVEL DE EDUCACIÓN DE SU PADRE, NIVEL DE EDUCACIÓN DE SU MADRE, ¿ENTRE QUÉ RANGO SE ENCUENTRA EL INGRESO, MENSUAL DE SU GRUPO FAMILIAR?, asigMasAtrasada sigla, asigMasAtrasada nombre, asigMasAtrasada periodoMalla, asigMasAtrasada periodoAvance, cantidadSemestresCronologicos, asigMasAtrasada cantidad, since most of the observations were null values. This reduced the dataset to 3,750 observations and 16 variables.

The variables of the final dataset can be summarized in: personal (ID ALUMNO, GENERO), location (COMUNA, REGIÓN), academic performance before university (TIPO COLEGIO, NEM, PUNTAJE RANKING, Nota4Medio), university application (COHORTE, TIPO POSTULACION, PREFERENCIA), university admission test scores (PSU MAT, PSU LENGUAJE, PSU OPCIONAL, PUNTAJE PONDERADO), and first-year performance (Semestres), being the last variable the class label.

After a conditional distribution analysis (details in [B](#)), we observed a lower value in the variables: NEM, PSU MAT, PSU OPCIONAL, PUNTAJE PONDERADO, PUNTAJE RANKING, and Nota4Medio increases the probability of dropout. This was expected since all these variables are related to the performance of the student. Moreover, students coming from public or private (with state support) schools have lower probabilities of dropout. This effect could be explained by the fact that Universidad Adolfo Ibáñez is an expensive private university, so students with lower resources probably entered on scholarships, which are granted to good students.

## 5.2 Universidad de Talca

Universidad de Talca data had 4 datasets: **BDD**, with 5,654 observations and 39 variables, **Notas 1er año**, with 59,578 observations and 7 variables, **Situaciones Especiales**, with 1,380 observations and 7 variables, and **Inscritas Reprobadas Por Año**, with 6,455 observations and 46 variables. Each observation corresponds to a student from Universidad de Talca. Even though we have a large quantity of data, the datasets contained several null values, and variables that did not contribute to the prediction of dropout, and were eliminated. Next, we described the data cleaning procedure, justifying the deletion of unnecessary variables and rows.

We analyzed if any dataset was not useful for first-year dropout prediction. We discarded the dataset **Notas 1er año** since first-year grades are gathered after the first year is completed. We also discarded the dataset **Situaciones Especiales**, which contains special situations for students, such as studying abroad for a semester. An analysis of this dataset shows us that only 43 observations were actually for first-year students, and they were not related to the problem. Apart from this, we noticed that the dataset **Inscritas Reprobadas Por Año**, which contains all the subjects taken and failed by the student by year, could help us to identify dropout from first-year students (the class label), since there is no explicit variable with this value. The dataset **BDD** was also maintained since it contains most of the variables that would be useful for predicting first-year dropout.

To create student's labels from the dataset **Inscritas Reprobadas Por Año**, we gathered all students that only took subjects during the first-year, and considered them as first-year dropout students. Specifically, we calculated the number of subjects taken for all the students from the second year to their last year. If the number of subjects taken by a student was 0, we assumed that the student dropout during the first year. The collected information was gathered in a new column called **Semestres** (for consistency with the dataset from UAI, subsection 5.1).

There were other options to create the same class label; however, those options showed important inconsistencies. The variables **ANHO\_INGRESO** and **ANHO\_ULT\_SIT** (entry year and last situation year respectively), from the dataset **BDD**, could be compared to check if they had the same values, meaning that the last situation of the student was the same year that the student signed in the university. On the other hand, the variable **DES\_A1**, also from **BDD**, was supposed to indicate if the student dropped out during the first year. However, these two methods did not include students who moved to another university during their first year, and, in some cases, their last situation year was inconsistent with the last semester in classes. Therefore, the most reliable method was to check if the student only had subjects enrolled during the first year.

At this point, we remained with 2 datasets: **Inscritas Reprobadas Por Año** (to generate the class label), and **BDD**. The class label from the dataset **Inscritas Reprobadas Por Año** was merged to the dataset **BDD** using the variable **Matricula** (unique ID for students), resulting in a unified dataset with 5,652 observations and 40 variables. However, it was still necessary to reduce unnecessary variables and observations.

We started by discarding five variables because of data quality:

- **Nacionalidad**: Nationality of the student. Most of their values are collected since 2018, and they have mainly a single value (Chile or no country).
- **Codigo Carrera**: Career code. Has duplicated information from the variable **Carrera**.
- **Anho Ingreso**: Entry year of the student. Has duplicated information from the variable **Anho**.
- **PROM PSU**: Mean of the variables **PTJE LENG** and **PTJE MAT**. It was used to fix incorrect values from **PTJE MAT**. Given the direct relation between **PTJE LENG** and **PTJE MAT**, this variable provokes collinearity, affecting the performance of machine learning models.

- **NEM:** Another scale for the variable PTJE NEM (with small variations, due to scale adjustments year by year). Similar to PROM PSU, this variable produces collinearity, affecting the performance of machine learning models.

This process reduced the dataset to 5,652 observations and 35 variables. A second set of variables were eliminated because their information is gathered after the first year is completed, therefore not useful for first-year dropout prediction:

- **Anho Egreso:** Graduation year of the student.
- **Grupo Año Egreso:** Graduation year group of the student.
- **Año egreso o titulación:** Graduation year of the student. Also, it was a duplicated variable of Anho Egreso.
- **Semestre egreso o titulación:** Semester (first or second of the year) of graduation.
- **Num. Asig. Repr. 1er año:** Failed subjects in first-year.
- **Situación al 2018:** Student situation at 2018 (graduated or dropped out)

These actions reduced the observations to 5,652 and 29 variables. However, more extensive analysis and actions were applied to increase the data quality.

Figure 5.1 shows the null values by year, in the variables related to PTJE CS ESPECÍFICA and PTJE HISTORIA. When observed by year (variable Anho), it was found that almost all of the null values were concentrated in the variables PTJE CS ESPECÍFICA QUÍMICA, PTJE CS ESPECÍFICA MAT, PTJE CS ESPECÍFICA FÍSICA, and PTJE CS ESPECÍFICA CIENCIAS SOCIALES. While the variable PTJE CS ESPECÍFICA BIOLOGÍA, is almost null in the year 2013, the variable PTJE HISTORIA is null in nearly half of the observations, which was expected considering that it is an optional test. All of these variables, with the exception of PTJE HISTORIA, were combined in a single variable called PTJE CS, containing the maximum value between them (if two or more values were available).

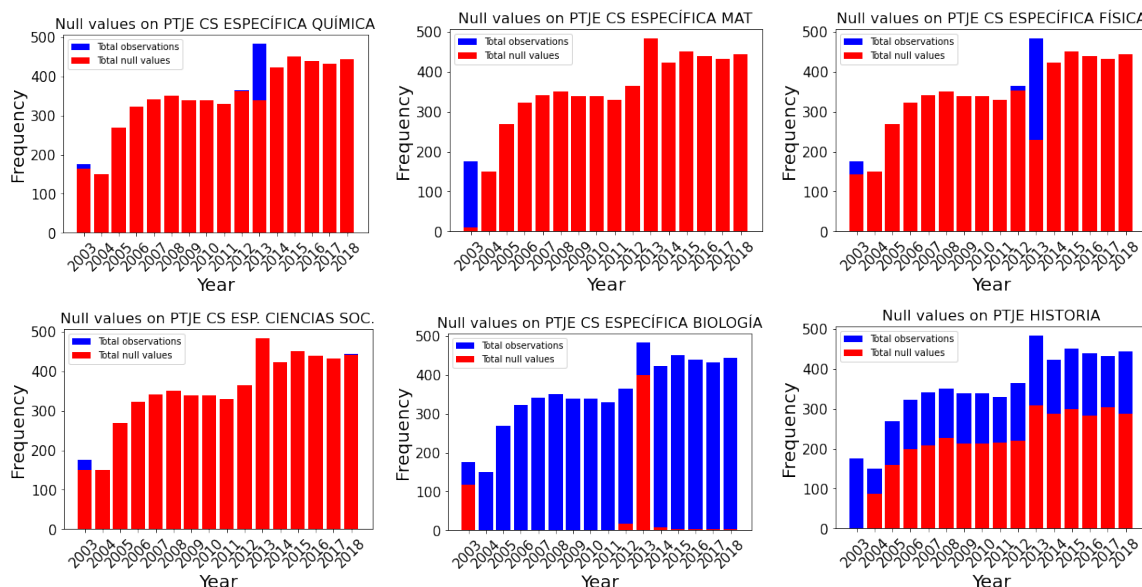


FIGURE 5.1: Null values by variable Anho in optional PSU tests.

As can be observed in figure 5.2, the variable PTJE RKG is null in every observation before 2013, as this score was created in that year. Given the importance of this variable, since it is the only measure of student performance related to behavior in their school, the observations before 2013



were eliminated. Also, the observations after 2018 were eliminated, because we do not have the status of the student after the first year.

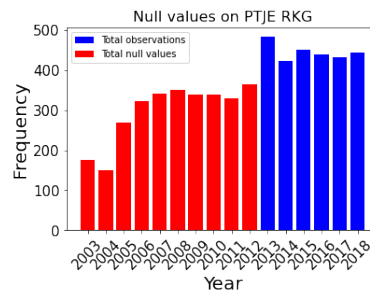


FIGURE 5.2: Null values by variable Anho in PTJE RKG.

Most variables related to the location of the students had many possible values (the majority with a low frequency). The following values were grouped to facilitate analysis:

- **Domicilio Ciudad:** City where the student comes from. It was reduced from its 56 possible values to 6: the 5 most common values, and Others, which gather all other cities.
- **Domicilio Comuna:** Commune where the student comes from. It was reduced from its 129 possible values to 10: the 9 most common values, and Others, which gather all other communes.
- **Domicilio Provincia:** Province where the student comes from. It was reduced from its 39 possible values to 6: the 5 most common values, and Others, which gather all other provinces.
- **Domicilio Region:** Region where the student comes from. It was reduced from its 13 possible values to 3: 6, 7 and Others, which gather all other regions.

We continue our analysis with the variable Carrera, where all the INGENIERÍA prefixes were removed for simplicity. Also, there were duplicated values with different names, so they were modified since they have the same meaning. First, EN CONSTRUCCION was replaced with CIVIL EN OBRAS CIVILES. Second, CIVIL EN MECATRÓNICA and EN MECATRÓNICA were replaced with CIVIL MECATRÓNICA. Third, MECANICA was replaced with CIVIL MECÁNICA. Finally, EN BIOINFORMÁTICA was replaced with CIVIL EN BIOINFORMÁTICA.

In our analysis, we also figure out a small problem with PTJE PPS, where 3 wrong values were noticed. Given the importance of PPS, the low number of data points, and the uncertainty in their possible prediction, we prefer to remove them.

Finally, the variable Fecha Nac (year of birth) was deleted since it does not influence the dropout decision, reducing the dataset to 2,201 observations and 24 variables.

The variables of the final dataset can be summarized in: personal (Rut, Sexo, Matricula, Ingreso bruto familiar, Educ padres), location (Domicilio Ciudad, Domicilio Comuna, Domicilio Provincia, Domicilio Region, Campus), academic performance before university (Tipo Educacion, Dependencia, PTJE NEM, PTJE RKG), university application (Anho, Grupo Ingreso, Anho Ingreso, Preferencia), university admission test scores (PTJE PPS, PTJE LENG, PTJE MAT, PTJE CS, PTJE HISTORIA), and first-year performance (Semestres), being this last variable the class label.

After a conditional distribution analysis (details in [B](#)), we observed that lower values in variables PTJE PPS, PTJE MAT, PTJE CS, PTJE NEM, and PTJE RKG increase the probability of dropout. This could be expected since all these variables are related to the performance of the student. We could also observe that lower family incomes and not-professional parents increase the probability

of dropout. It is also important to note that the selected career also affects dropout probability. Specifically, CIVIL EN COMPUTACIÓN, CIVIL DE MINAS, or CIVIL EN BIOINFORMÁTICA has higher dropouts than other degrees.

### 5.3 Unification of both datasets

After the analysis of the datasets of Universidad Adolfo Ibáñez and Universidad de Talca, we unified them, creating a new dataset containing the shared variables and renaming them for simplicity. This new dataset contains the following variables:

- **ID**: From ID ALUMNO and Matricula.
- **year**: From COHORTE and Anho.
- **gender**: From GENERO and Sexo.
- **school**: From TIPO COLEGIO and Dependencia.
- **admission**: From TIPO POSTULACION and Grupo Ingreso.
- **nem**: From NEM and PTJE NEM.
- **ranking**: From PUNTAJE RANKING and PTJE RKG.
- **mat**: From PSU MAT and PTJE MAT.
- **lang**: From PSU LENGUAJE and PTJE LENG.
- **optional**: From PSU OPCIONAL and the maximum value between PTJE HISTORIA and PTJE CS.
- **pps**: From PUNTAJE PONDERADO and PTJE PPS
- **preference**: From PREFERENCIA and Preferencia.
- **commune**: From COMUNA and Domicilio Comuna.
- **region**: From REGIÓN and Domicilio Region.
- **semestres**: From Semestres in both datasets, being the class label.
- **university**: Containing the university of the student (UAI or Talca). Created for identification of the observations after unification of the datasets.

It was also necessary to rename the values in some variables since they have the same meaning. Specifically, this required the renaming of the values in the variables school and admission. There was a special problem with the variables commune and region since their values were grouped with different categories. So, it was necessary to regroup them obtaining the same categories values among both datasets.

This new dataset ended with 5,951 observations and 16 variables. It is important to notice that there are more observations from Universidad Adolfo Ibáñez (3,750 observations), which must be handled in the application of machine learning models. Also, the conditional distribution analysis was not done in this dataset since it would be very similar to the analysis of each one separately.



## Chapter 6

# Results and discussion

In this section, we show and discuss the final results of each model, where a variable and parameter selection was applied. All results correspond to the  $F_1$  score (positive and negative), accuracy, and AUC of 10-fold cross-validation of the best tuned model. The variable selection was done through forward selection (see section 4), and the hyperparameters of each model were searched through potential combinations of their parameters. For a detailed explanation of the tuning process for each model refer to C.

We applied the models: KNN, SVM, decision tree, random forest, gradient boosting decision tree, naive Bayes, logistic regression, and neural network. We included a random model to compare if the models behave better than an aleatory decision. The results for both-universities-datasets (universidad Adolfo Ibáñez dataset and Universidad de Talca) are shown below.

### 6.1 Both-universities-dataset results

The results for both-universities-dataset are shown in table 6.1. A hypothesis testing comparing if the scores were significantly different was also applied ( $\alpha = 0.05$ ). Finally, the hyperparameters tuning process obtained the following parameters:

- **KNN:**  $K = 29$ .
- **SVM:**  $C = 10$ , polynomial kernel.
- **Decision tree:** minimum samples required to be a leaf node of 187.
- **Random forest:** minimum samples required to be a leaf node of 100, 500 trees, 20 features per tree.
- **Gradient boosting decision tree:** minimum samples required to be a leaf node of 150, 110 trees, 6 features per tree
- **Naive Bayes:** Gaussian distribution were assumed.
- **Logistic regression:** Only variable selection was applied.
- **Neural Network:** 2 hidden layers, 15 neurons per layer.

The  $F_1$  score + class showed that all models have similar performance, with SVM and gradient boosting being significantly better than decision tree and naive Bayes. The  $F_1$  score - class and accuracy showed that random forest and gradient boosting are significantly better than the other models, with exception of decision tree. It also showed that logistic regression and KNN are significantly worse than the other models. Finally, the AUC showed that all models have similar scores. It is also important to note that all models behaved better than the random model.

TABLE 6.1: Models results. Best scores for each model. Both universities.

| Model               | $F_1$ score + class | $F_1$ score - class | Accuracy        | AUC             |
|---------------------|---------------------|---------------------|-----------------|-----------------|
| Random model        | $0.27 \pm 0.02$     | $0.63 \pm 0.02$     | $0.51 \pm 0.02$ | $0.50 \pm 0.04$ |
| KNN                 | $0.35 \pm 0.03$     | $0.73 \pm 0.02$     | $0.62 \pm 0.02$ | $0.63 \pm 0.03$ |
| SVM                 | $0.36 \pm 0.02$     | $0.76 \pm 0.02$     | $0.65 \pm 0.02$ | $0.65 \pm 0.03$ |
| Decision tree       | $0.33 \pm 0.03$     | $0.79 \pm 0.03$     | $0.68 \pm 0.03$ | $0.64 \pm 0.03$ |
| Random forest       | $0.35 \pm 0.03$     | $0.80 \pm 0.02$     | $0.69 \pm 0.02$ | $0.65 \pm 0.03$ |
| Gradient boosting   | $0.37 \pm 0.03$     | $0.80 \pm 0.01$     | $0.69 \pm 0.02$ | $0.66 \pm 0.03$ |
| Naive Bayes         | $0.34 \pm 0.02$     | $0.77 \pm 0.01$     | $0.66 \pm 0.01$ | $0.64 \pm 0.02$ |
| Logistic regression | $0.35 \pm 0.03$     | $0.73 \pm 0.01$     | $0.62 \pm 0.02$ | $0.65 \pm 0.03$ |
| Neural network      | $0.35 \pm 0.03$     | $0.76 \pm 0.03$     | $0.66 \pm 0.03$ | $0.65 \pm 0.02$ |

In addition to the performance, the influence of individual variables was examined through the learned parameters of the decision tree, random forest, gradient boosting, naive Bayes, and logistic regression models. The decision tree, dropping some variables to avoid collinearity, showed that pps, mat, and lang are the most important variables, being pps the root variable in most of the folds. Figure 6.1 shows the learned tree for one of the folds of cross-validation, and the average feature importance for each variable, showing significant importance to pps, mat, and lang, in the same way as the tree.

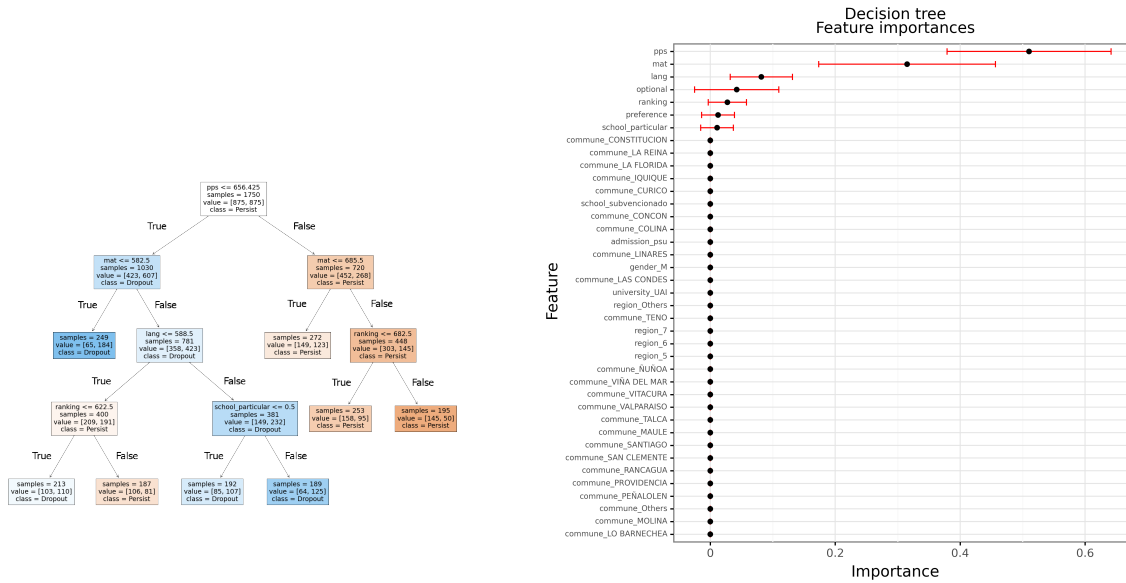


FIGURE 6.1: Decision tree learned tree of one fold, and average feature importance. Both universities.

The random forest features importance for one fold, dropping some variables to avoid collinearity, are shown in figure 6.2, where we can see that mat, pps, and optional are the most important variables. We can also see that only 5 variables (mat, pps, optional, ranking, and lang) have significant importance.

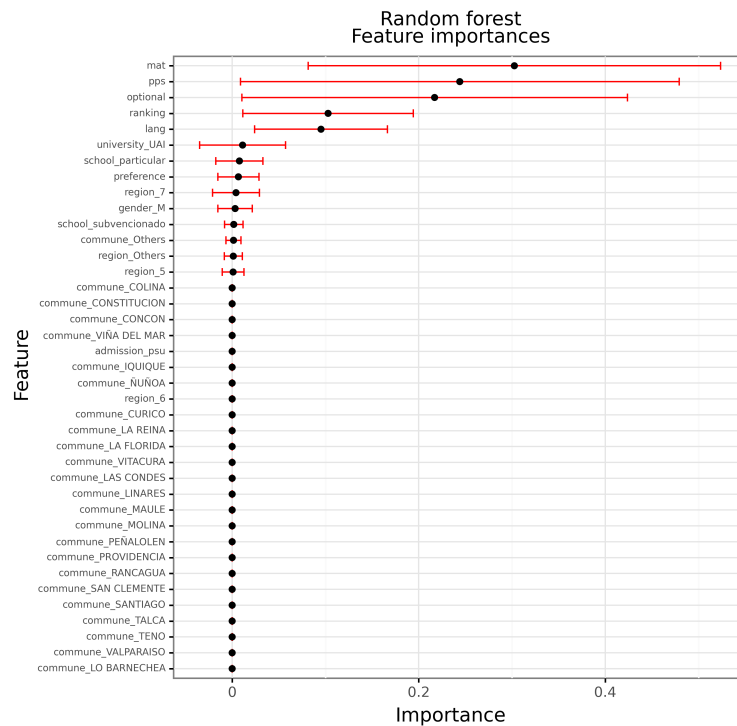


FIGURE 6.2: Random forest feature importance of one fold. Both universities.

The gradient boosting decision tree average features importance, dropping some variables to avoid collinearity, are shown in figure 6.3. The results show higher importance to *mat*, *pps*, *optional*, *ranking*, and *lang*. Note that these are similar to the previous tree-based models (decision tree and random forest) selected variables.

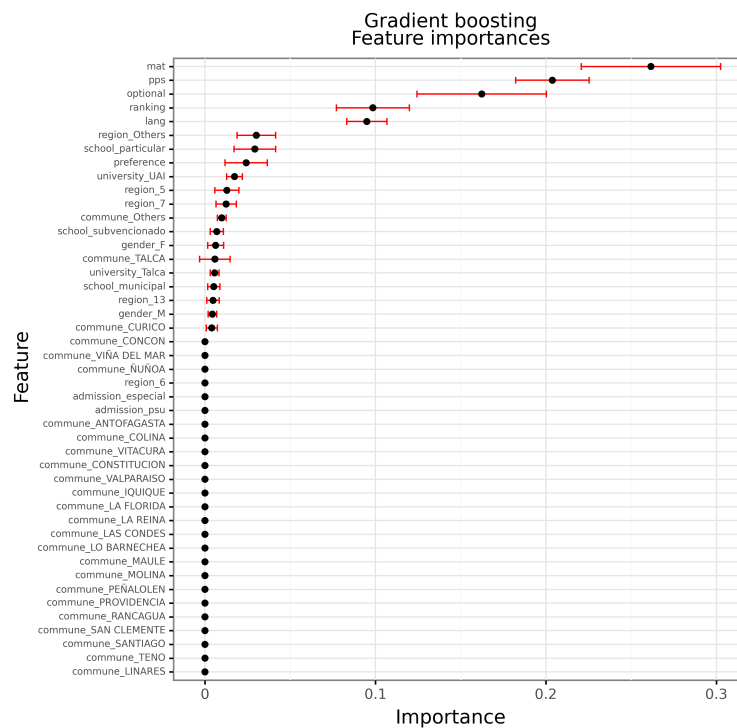


FIGURE 6.3: Gradient boosting average feature importance. Both universities.

For naive Bayes, table 6.2 shows the means  $\pm$  one standard deviation of the learned distributions per class of one fold, and the p-value for the hypothesis test comparing the difference between means. The probability distributions show that Gaussian means are significantly higher for the negative class in all variables (mat, optional, ranking, pps, and nem). This implies a lower probability of dropout when any of these values increase.

TABLE 6.2: Naive Bayes distributions of one fold. Both universities.

| Variable | mean - class       | mean + class       | p-value                | n observations |
|----------|--------------------|--------------------|------------------------|----------------|
| mat      | 669.19 $\pm$ 61.26 | 639.24 $\pm$ 66.17 | $2.23 \times 10^{-21}$ | 840            |
| optional | 620.58 $\pm$ 62.24 | 595.34 $\pm$ 68.84 | $5.72 \times 10^{-15}$ | 840            |
| ranking  | 664.75 $\pm$ 86.74 | 641.36 $\pm$ 94.45 | $1.42 \times 10^{-07}$ | 840            |
| pps      | 657.76 $\pm$ 47.00 | 634.97 $\pm$ 51.31 | $7.47 \times 10^{-21}$ | 840            |
| nem      | 639.23 $\pm$ 67.51 | 617.37 $\pm$ 72.67 | $2.17 \times 10^{-10}$ | 840            |

The learned parameters of the logistic regression, standardized model and dropping some variables to avoid collinearity, are shown in Table 6.3. The results show higher importance to the variables mat, pps, optional, and lang, where a higher mat, pps, and pps scores imply a lower probability of dropout. In contrast, a higher lang score implies a higher probability of dropout which may be because engineering students may not be the best in language. Table 6.3 also includes the p-value of a hypothesis testing (if the learned parameter is significantly different than 0), showing that all the included variables with exception of school\_particular had significant importance.

TABLE 6.3: Logistic regression average  $\beta$  value for each variable. Both universities.

| Variable             | $\beta$ value    | p-value                |
|----------------------|------------------|------------------------|
| mat                  | -2.28 $\pm$ 0.27 | $1.25 \times 10^{-15}$ |
| pps                  | -1.54 $\pm$ 0.19 | $2.95 \times 10^{-15}$ |
| optional             | -1.44 $\pm$ 0.34 | $7.48 \times 10^{-11}$ |
| gender_M             | -0.32 $\pm$ 0.24 | $4.66 \times 10^{-04}$ |
| school_particular    | 0.04 $\pm$ 0.09  | 0.19                   |
| school_subvencionado | 0.17 $\pm$ 0.09  | $1.41 \times 10^{-05}$ |
| admission_psu        | -0.37 $\pm$ 0.08 | $1.28 \times 10^{-11}$ |
| lang                 | 1.23 $\pm$ 0.26  | $1.00 \times 10^{-11}$ |

The results for both-universities-dataset showed that, as an initial conclusion, it is possible to train a model that has better performance than an aleatory decision. However, all of the models showed similar performance, so it is not possible to select any of them as the best. Despite that, the interpretative models (decision tree, random forest, naive Bayes, and logistic regression) showed higher importance for the variables mat, pps, optional, and lang. The variables pps and mat were considered relevant in all these models, while optional and lang only in 4 of them. A higher score in mat, pps, and optional shows a reduction in the dropout probability. However, surprisingly, a higher score in lang increases the probability of dropout, which could be explained by the fact that we analyze engineering students. Finally, it is important to note that all the common relevant variables are numerical, which may imply that the categorical data is less relevant.

## 6.2 Universidad Adolfo Ibáñez results

The results for Universidad Adolfo Ibáñez are shown in table 6.4. A hypothesis testing comparing if the scores were significantly different was also applied ( $\alpha = 0.05$ ). Finally, the hyperparameters

tuning process obtained the following parameters:

- **KNN:**  $K = 17$ .
- **SVM:**  $C = 1$ , polynomial kernel.
- **Decision tree:** minimum samples required to be a leaf node of 48.
- **Random forest:** minimum samples required to be a leaf node of 20, 50 trees, 15 features per tree.
- **Gradient boosting decision tree:** minimum samples required to be a leaf node of 60, 80 trees, 20 features per tree.
- **Naive Bayes:** Gaussian distribution was assumed.
- **Logistic regression:** Only variable selection was applied.
- **Neural Network:** 1 hidden layer, with 18 neurons.

The  $F_1$  score + class showed that all models have similar performance. The  $F_1$  score - class and accuracy showed that random forest and decision trees are significantly better than the other models, being the first significantly better than the second. Finally, the AUC showed that all models have similar scores.

TABLE 6.4: Models results. Best scores for each model. UAI.

| Model               | $F_1$ score + class | $F_1$ score - class | Accuracy        | AUC             |
|---------------------|---------------------|---------------------|-----------------|-----------------|
| Random model        | $0.26 \pm 0.03$     | $0.64 \pm 0.01$     | $0.51 \pm 0.01$ | $0.51 \pm 0.03$ |
| KNN                 | $0.30 \pm 0.05$     | $0.72 \pm 0.02$     | $0.60 \pm 0.02$ | $0.61 \pm 0.04$ |
| SVM                 | $0.31 \pm 0.05$     | $0.69 \pm 0.04$     | $0.57 \pm 0.05$ | $0.62 \pm 0.06$ |
| Decision tree       | $0.28 \pm 0.03$     | $0.78 \pm 0.04$     | $0.66 \pm 0.05$ | $0.61 \pm 0.05$ |
| Random forest       | $0.30 \pm 0.06$     | $0.82 \pm 0.01$     | $0.71 \pm 0.02$ | $0.62 \pm 0.05$ |
| Gradient boosting   | $0.31 \pm 0.04$     | $0.73 \pm 0.02$     | $0.61 \pm 0.03$ | $0.63 \pm 0.03$ |
| Naive Bayes         | $0.29 \pm 0.04$     | $0.68 \pm 0.03$     | $0.56 \pm 0.04$ | $0.63 \pm 0.04$ |
| Logistic regression | $0.30 \pm 0.05$     | $0.72 \pm 0.03$     | $0.60 \pm 0.03$ | $0.62 \pm 0.05$ |
| Neural network      | $0.28 \pm 0.02$     | $0.67 \pm 0.10$     | $0.55 \pm 0.10$ | $0.63 \pm 0.03$ |

As with the both-universities-dataset, the influence of individual variables was examined through the learned parameters of the decision tree, random forest, gradient boosting, naive Bayes, and logistic regression models. The decision tree, dropping some variables to avoid collinearity, showed that pps, lang, and mat are the most important variables, being pps the root variable in most of the folds (note that these are the same important variables as in the both-universities-dataset). Figure 6.4 shows the learned tree for one of the folds of cross-validation, and the average feature importance for each variable, showing significant importance to pps, lang, optional, mat, and ranking.

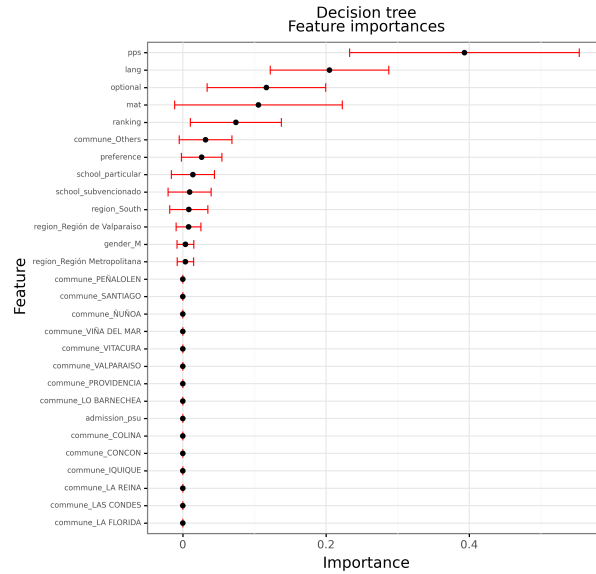


FIGURE 6.4: Decision tree learned tree of one fold, and average feature importance. Uai.

The random forest features importance for one fold, dropping some variables to avoid collinearity, are shown in figure 6.5, where we can see that pps, lang, optional, ranking, and mat are the most important variables. Note that similar to the decision tree model, these are the same important variables as in the both-universities-dataset, but in different order of importance.

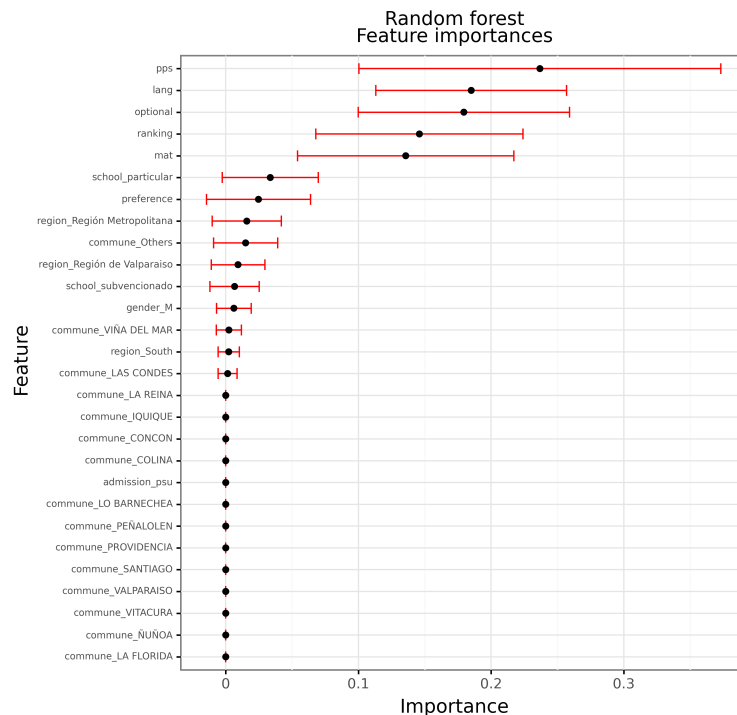


FIGURE 6.5: Random forest feature importance of one fold. Uai.

The gradient boosting decision tree average feature importance of each variable, dropping some variables to avoid collinearity, is shown in figure 6.6. The results show higher importance to pps, optional, lang, ranking, and mat. As in the first dataset, note that these are similar to the previous

tree-based models (decision tree and random forest) selected important variables. Further that, it considered the same important variables as the both-universities-dataset gradient boosting, but with different importance.

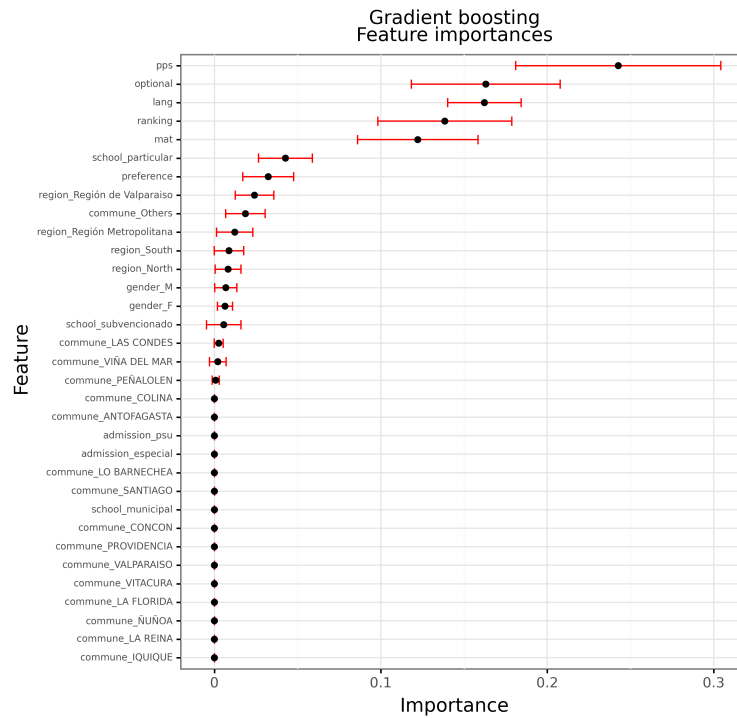


FIGURE 6.6: Gradient boosting average feature importance. Uai.

For naive Bayes, table 6.5 shows the means  $\pm$  one standard deviation of the learned distributions per class of one fold, and the p-value for the hypothesis test comparing the difference between means. The probability distributions show that Gaussian mean is significantly higher for the negative class in the variables mat, optional, and ranking. This implies a lower probability of dropout when any of these values increase.

TABLE 6.5: Naive Bayes distributions of one fold. Uai.

| Variable | mean - class       | mean + class       | p-value                | n observations |
|----------|--------------------|--------------------|------------------------|----------------|
| mat      | 697.58 $\pm$ 50.0  | 684.2 $\pm$ 47.51  | 6.85 $\times 10^{-05}$ | 425            |
| optional | 642.40 $\pm$ 53.86 | 630.35 $\pm$ 60.16 | 0.002                  | 425            |
| lang     | 615.74 $\pm$ 69.14 | 620.31 $\pm$ 67.12 | 0.329                  | 425            |
| ranking  | 656.44 $\pm$ 88.0  | 628.67 $\pm$ 89.79 | 6.07 $\times 10^{-06}$ | 425            |

The learned parameters of the logistic regression, standardized model and dropping some variables to avoid collinearity, are shown in Table 6.6. The results show higher importance to the variables pps, lang, and region, where a higher pps score, and being from region Metropolitana, from Valparaíso, or the south implies a lower probability of dropout, where last of them has almost no importance. In contrast, a higher lang score implies a higher probability of dropout, this, as stated with the both-universities-dataset, could be because engineering students may not be the best in language. Note that, surprisingly, this model includes a categorical variable as one of the most important ones, unlike the previous ones. Table 6.6 also includes the p-value of a hypothesis testing (if the learned parameter is significantly different than 0), showing that all the included variables had significant importance.

TABLE 6.6: Logistic regression average  $\beta$  value for each variable. Uai.

| Variable                    | $\beta$ value    | p-value                |
|-----------------------------|------------------|------------------------|
| pps                         | $-2.46 \pm 0.39$ | $1.53 \times 10^{-13}$ |
| region_Región de Valparaíso | $-0.67 \pm 0.19$ | $2.25 \times 10^{-09}$ |
| region_Región Metropolitana | $-0.43 \pm 0.15$ | $3.34 \times 10^{-08}$ |
| region_South                | $-0.12 \pm 0.12$ | 0.004                  |
| gender_M                    | $0.17 \pm 0.15$  | 0.002                  |
| lang                        | $1.26 \pm 0.38$  | $4.14 \times 10^{-09}$ |

The results for Universidad Adolfo Ibáñez showed that, as an initial conclusion, it is possible to train a model that has better performance than an aleatory decision. When considering the  $F_1$  score - class, or the accuracy, it is possible to select random forest and decision tree as the best models. The interpretative models (decision tree, random forest, gradient boosting, naive Bayes, and logistic regression) showed higher importance for the variables mat, pps, and lang. These variables were considered relevant in 4 of these models, but, unlike in the both-universities-dataset, there are no variables considered relevant in all the models. A higher score in mat and pps shows a reduction in the dropout probability. However, as with the both-universities-dataset, a higher score in lang increases the probability of dropout, which could be explained by the fact that we analyze engineering students. Finally, it is important to note that, again, all the common relevant variables are numerical, which may imply that the categorical data is less relevant (except for region).

### 6.3 Universidad de Talca results

The results for Universidad de Talca are shown in table 6.7. A hypothesis testing comparing if the scores were significantly different was also applied ( $\alpha = 0.05$ ). Finally, the hyperparameters tuning process obtained the following parameters:

- **KNN:**  $K = 71$ .
- **SVM:**  $C = 1$ , polynomial kernel.
- **Decision tree:** minimum samples required to be a leaf node of 123.
- **Random forest:** minimum samples required to be a leaf node of 150, 50 trees, 15 features per tree.
- **Gradient boosting decision tree:** minimum samples required to be a leaf node of 130, 50 trees, 15 features per tree, maximum depth of 10.
- **Naive Bayes:** Gaussian distribution was assumed.
- **Logistic regression:** Only variable selection was applied.
- **Neural Network:** 1 hidden layer, with 18 neurons.

The  $F_1$  score + class showed that all the models have similar performance, with no significant difference between any of them. The  $F_1$  score - class showed something completely different, it showed that KNN is significantly better than all the models, with exception of random forest and neural network. Note that the neural network model showed a higher deviation. In the accuracy, KNN again was a good model, but only significantly better than SVM, decision tree, and logistic regression. Note that again, the neural network model showed a higher deviation. Finally, the AUC showed, as the first measure, no significant difference between any model.



TABLE 6.7: Models results. Best scores for each model. Universidad de Talca.

| Model               | $F_1$ score + class | $F_1$ score - class | Accuracy        | AUC             |
|---------------------|---------------------|---------------------|-----------------|-----------------|
| Random model        | $0.31 \pm 0.04$     | $0.63 \pm 0.04$     | $0.52 \pm 0.04$ | $0.52 \pm 0.06$ |
| KNN                 | $0.42 \pm 0.05$     | $0.76 \pm 0.02$     | $0.66 \pm 0.03$ | $0.66 \pm 0.03$ |
| SVM                 | $0.42 \pm 0.03$     | $0.71 \pm 0.03$     | $0.61 \pm 0.03$ | $0.67 \pm 0.02$ |
| Decision tree       | $0.41 \pm 0.05$     | $0.73 \pm 0.03$     | $0.63 \pm 0.03$ | $0.66 \pm 0.03$ |
| Random forest       | $0.41 \pm 0.05$     | $0.74 \pm 0.03$     | $0.64 \pm 0.03$ | $0.66 \pm 0.03$ |
| Gradient boosting   | $0.41 \pm 0.05$     | $0.73 \pm 0.04$     | $0.63 \pm 0.04$ | $0.65 \pm 0.04$ |
| Naive Bayes         | $0.42 \pm 0.03$     | $0.74 \pm 0.02$     | $0.64 \pm 0.03$ | $0.66 \pm 0.02$ |
| Logistic regression | $0.41 \pm 0.03$     | $0.73 \pm 0.01$     | $0.63 \pm 0.02$ | $0.66 \pm 0.03$ |
| Neural network      | $0.39 \pm 0.05$     | $0.73 \pm 0.09$     | $0.63 \pm 0.08$ | $0.66 \pm 0.02$ |

The influence of individual variables through the learned parameters of decision tree, random forest, gradient boosting, naive Bayes, and logistic regression models was examined as in the previous datasets. The decision tree, dropping some variables to avoid collinearity, showed that mat, ranking, and lang are the most important variables, being mat the root variable in all of the folds (Figure 6.7, left plot). Figure 6.7 shows the learned tree for one of the folds of cross-validation, and the average feature importance for each variable, showing significant importance to ranking and lang and particularly high importance to the mat variable.

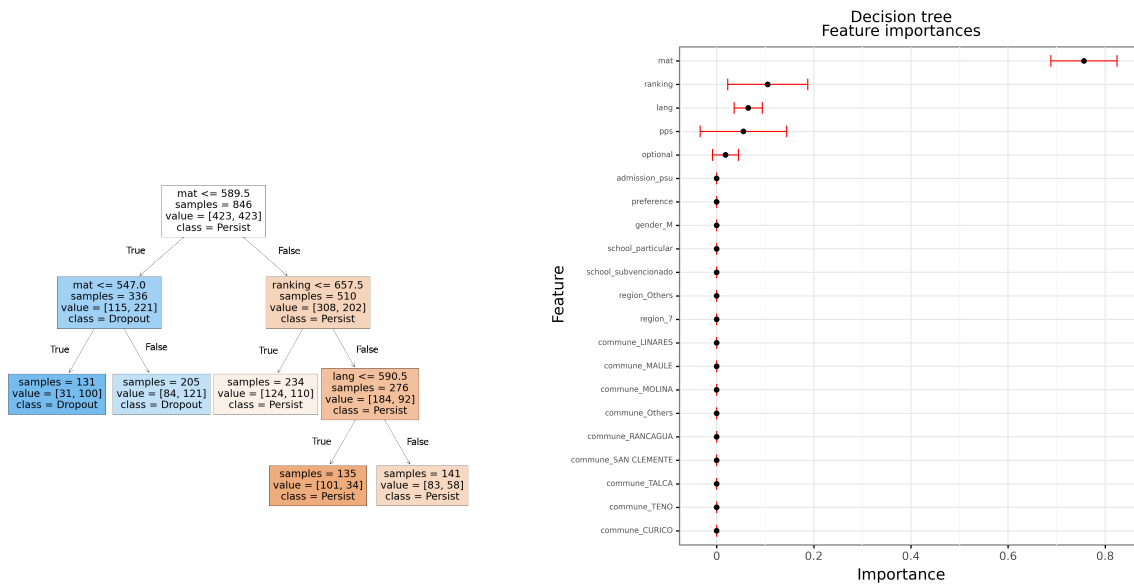


FIGURE 6.7: Decision tree learned tree of one fold, and average feature importance. U de Talca.

The random forest features importance for one fold, dropping some variables to avoid collinearity, are shown in figure 6.8, where we can see that only the mat variable is significantly important, similar to the fact that in the decision tree this was the most important variable. It surprises the fact that it only uses a single variable and gets a similar result to the other models.

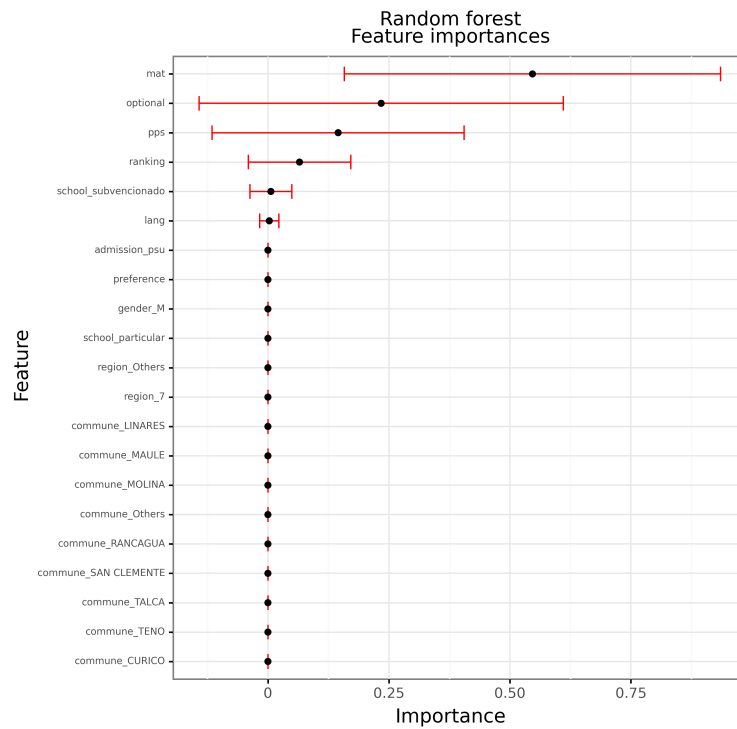


FIGURE 6.8: Random forest feature importance of one fold. U de Talca.

The gradient boosting decision tree average feature importance of each variable, dropping some variables to avoid collinearity, is shown in figure 6.9. The results show higher importance to *mat*, *optional*, *pps*, *lang*, and *ranking*, however, we can note that the importance for some of them is not statistically significant. As in the previous datasets, note that these are similar to the previous tree-based models (decision tree and random forest) selected important variables. Once again, it considered the same important variables as the both-universities-dataset and UAI gradient boosting, but with different importance.

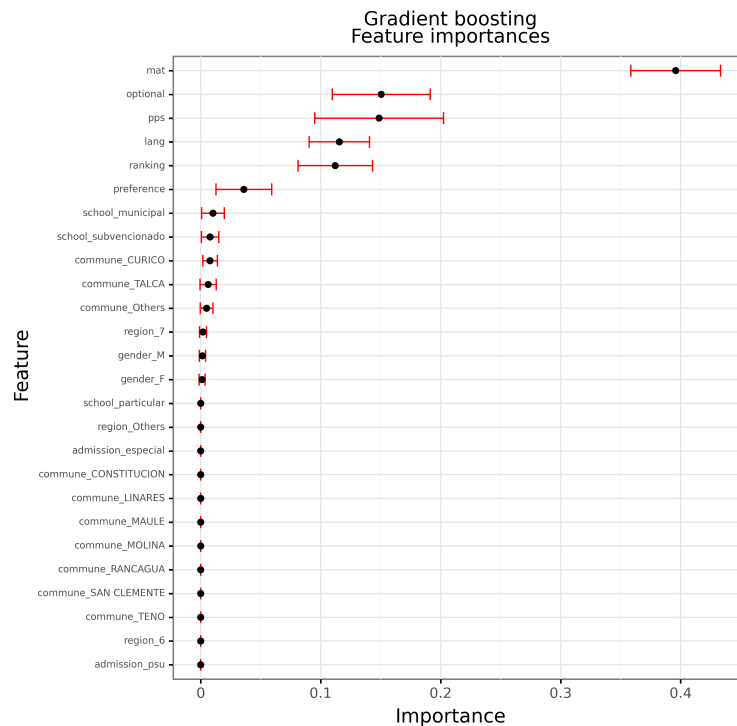


FIGURE 6.9: Gradient boosting average feature importance. U de Talca.

For naive Bayes, table 6.8 shows the means  $\pm$  one standard deviation of the learned distributions per class of one fold, and the p-value for the hypothesis test comparing the difference between means. The probability distributions show that Gaussian means are significantly higher for the negative class in both variables, implying a lower probability of dropout when any of these values increase. Note that naive Bayes differs from the tree-based models by including nem as an important variable, however, nem is highly correlated with ranking, which was included in decision tree.

TABLE 6.8: Naive Bayes distributions of one fold. U de Talca.

| Variable | mean - class       | mean + class       | p-value                | n observations |
|----------|--------------------|--------------------|------------------------|----------------|
| mat      | 622.87 $\pm$ 53.36 | 592.69 $\pm$ 55.66 | 2.43 $\times 10^{-15}$ | 412            |
| nem      | 642.47 $\pm$ 72.98 | 619.31 $\pm$ 74.53 | 5.44 $\times 10^{-06}$ | 412            |

The learned parameters of the logistic regression, standardized model and dropping some variables to avoid collinearity, are shown in Table 6.9. The results show higher importance to all the variables included, with a special high value in mat, as in the decision tree model. A higher mat, pps, and nem score, as well as a psu admission (regular admission or special admission, such as sport-scholarship), implies a lower probability of dropout. A higher lang score, on the other hand, implies a higher probability of dropout, a particular behavior also seeing in previous models in this thesis. A higher preference value (lower preference for Universidad de Talca) also increases this probability, which was expected. It surprises that this model includes more variables than the previous models, including even a categorical variable as one of the most important ones, uncommon in most of the analyses.

TABLE 6.9: Logistic regression average  $\beta$  value for each variable. U de Talca.

| Variable      | $\beta$ value    | p-value                |
|---------------|------------------|------------------------|
| mat           | $-2.72 \pm 0.23$ | $1.57 \times 10^{-18}$ |
| pps           | $-0.80 \pm 0.24$ | $4.18 \times 10^{-09}$ |
| admission_psu | $-0.57 \pm 0.27$ | $2.60 \times 10^{-06}$ |
| nem           | $-0.57 \pm 0.25$ | $1.19 \times 10^{-06}$ |
| lang          | $0.43 \pm 0.16$  | $1.20 \times 10^{-07}$ |
| preference    | $0.74 \pm 0.34$  | $1.79 \times 10^{-06}$ |

The results for Universidad de Talca showed that, as in the previous datasets, it is possible to train a model that has better performance than an aleatory decision. Even though the measures showed similar performance among the models, the higher average scores are in KNN, an interesting result considering that in the both-universities-dataset, KNN had the worst average score for the  $F_1$  score - class and the accuracy. Besides, the interpretative models (decision tree, random forest, naive Bayes, and logistic regression) showed higher importance for the variable mat. This variable was considered very important in all of the models, where a higher score shows a reduction in the dropout probability, as expected. Note that there are less important variables than in the previous analyses, this since most of the models considered different variables as important, with few of them being common, for example, the variables lang, nem, pps, and ranking were also considered important, but only in 2 models. Finally, once again all of the common relevant variables are numerical, which may imply that the categorical data is less relevant.

## 6.4 Universidad de Talca all variables

In section 5.3, the unification of the datasets discarded non-shared variables. This process eliminated the following variables from Universidad de Talca: Carrera (degree), Campus, Ingreso bruto familiar (family income), Educ padres (professional parents). In contrast, no variables were discarded from Universidad Adolfo Ibáñez. To analyze the difference between the unified dataset and the original dataset, we applied a new analysis for Universidad de Talca, including all these variables. Since all of the variables were categorical or ordinal, KNN and SVM were not analyzed. Additionally, the tuning process of Naive Bayes showed the same results as the previous analysis (section 6.3), so it is also discarded from this analysis. The results for all other models are shown in table 6.10, and the hyperparameter tuning process selected the following parameters:

- **Decision tree:** minimum samples required to be a leaf node of 102.
- **Random forest:** minimum samples required to be a leaf node of 2, 500 trees, 4 features per tree.
- **Gradient boosting decision tree:** minimum samples required to be a leaf node of 135, 180 trees, 5 features per tree, maximum depth of 4.
- **Logistic regression:** Only variable selection was applied.
- **Neural Network:** 1 hidden layer, with 5 neurons.

The  $F_1$  score + class,  $F_1$  score - class, and accuracy showed that all models have similar performance, while the AUC showed that the neural network is better than decision tree, random forest, and gradient boosting. The results also showed similar performance compared to the first Universidad de Talca analysis, with an apparent improvement in the  $F_1$  score + class and AUC in the decision tree. However, this difference was not significant.

TABLE 6.10: Models results. Best scores for each model. Universidad de Talca all variables.

| Model               | $F_1$ score + class | $F_1$ score - class | Accuracy        | AUC             |
|---------------------|---------------------|---------------------|-----------------|-----------------|
| Random model        | $0.31 \pm 0.04$     | $0.63 \pm 0.04$     | $0.52 \pm 0.04$ | $0.52 \pm 0.06$ |
| Decision tree       | $0.41 \pm 0.05$     | $0.72 \pm 0.04$     | $0.63 \pm 0.05$ | $0.65 \pm 0.03$ |
| Random forest       | $0.40 \pm 0.04$     | $0.72 \pm 0.03$     | $0.62 \pm 0.04$ | $0.66 \pm 0.03$ |
| Gradient boosting   | $0.40 \pm 0.04$     | $0.73 \pm 0.03$     | $0.63 \pm 0.03$ | $0.66 \pm 0.02$ |
| Logistic regression | $0.43 \pm 0.04$     | $0.74 \pm 0.03$     | $0.64 \pm 0.03$ | $0.68 \pm 0.03$ |
| Neural network      | $0.42 \pm 0.04$     | $0.72 \pm 0.08$     | $0.63 \pm 0.07$ | $0.69 \pm 0.03$ |

We examined the learned parameters of decision tree, random forest, gradient boosting, and logistic regression models, focusing on the non-shared variables. The decision tree and random forest models considered only shared variables. However, the gradient boosting and logistic regression models included non-shared variables as important features. Figure 6.10 and table 6.11 shows the gradient boosting average feature importance of each variable, and the average  $\beta$  values of the logistic regression, respectively. The gradient boosting showed a low (but significant) importance to degree civil Industrial and family income, while the logistic regression showed importance to the following degrees: civil mecánica, civil en obras civiles, and civil en bioinformática, where belonging to mecánica or obras civiles decrease the probability of dropout, while belonging to bioinformática increases it. Even though new non-shared variables were included, the effects of them in terms of  $F_1$ -score and accuracy are not statistically significant.

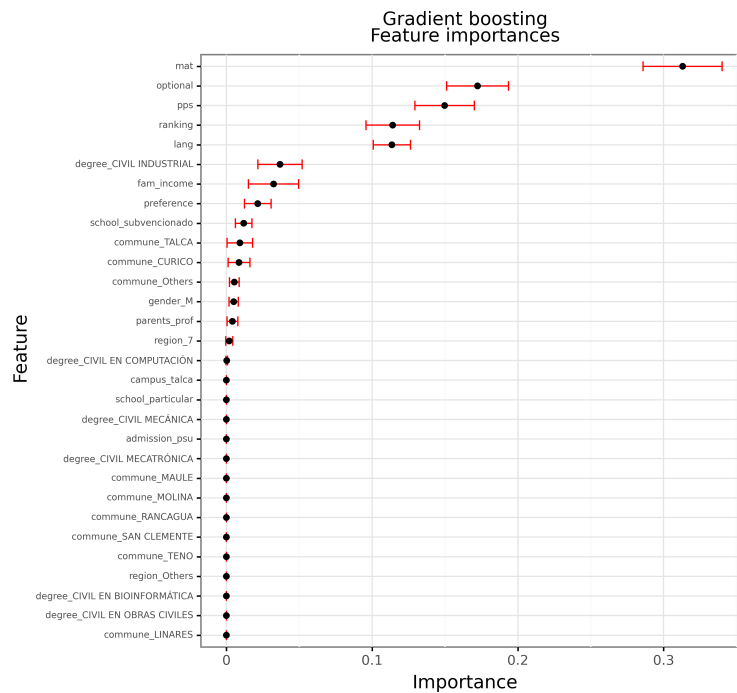


FIGURE 6.10: Gradient boosting average feature importance. U de Talca all variables.

TABLE 6.11: Logistic regression average  $\beta$  value for each variable. U Talca all variables.

| Variable                       | $\beta$ value    | p-value                |
|--------------------------------|------------------|------------------------|
| mat                            | $-2.94 \pm 0.28$ | $8.91 \times 10^{-18}$ |
| nem                            | $-0.85 \pm 0.23$ | $8.4 \times 10^{-10}$  |
| degree_CIVIL MECÁNICA          | $-0.52 \pm 0.12$ | $1.3e \times 10^{-07}$ |
| degree_CIVIL INDUSTRIAL        | $-0.48 \pm 0.11$ | 0.431                  |
| degree_CIVIL EN OBRAS CIVILES  | $-0.41 \pm 0.16$ | $2.02 \times 10^{-07}$ |
| degree_CIVIL MECATRÓNICA       | $-0.09 \pm 0.16$ | 0.072                  |
| degree_CIVIL EN COMPUTACIÓN    | $-0.02 \pm 0.12$ | 0.701                  |
| degree_CIVIL EN BIOINFORMÁTICA | $0.45 \pm 0.28$  | $1.21 \times 10^{-05}$ |

## 6.5 Discussion

The results from previous analyses are summarized in tables 6.12 to 6.15, where each table shows the results for one metric over all datasets (both-universities-dataset, Universidad Adolfo Ibáñez, Universidad de Talca, and Universidad de Talca All variables). In all the tables, “-” means that U Talca All Vars used the same variable as U Talca. Table 6.16 shows all variables that were considered important by at least one model, on any dataset. The notation is based on “Y” and “N” values, indicating if the variable was considered important by the model or not, while “-” means that the variable did not exist on that dataset. To summarize all datasets, the display of the values has the following pattern: “both,UAI,UTalca,Utalca All vars”.

Table 6.12 shows the  $F_1$  score + class of the models (since the data was highly unbalanced). Based on the results, It was not possible to select a single model as the best for all datasets. The best model could be gradient boosting, which had the higher average score in 2 of the 4 datasets, but they were not significantly better. On the contrary, based exclusively on the score, we could discard decision tree, since it had the lowest score in 2 datasets, and did not excel in any dataset. When comparing the performance per dataset, Universidad de Talca datasets have higher scores for each model. This may imply a better data quality for this university, but also could be due to their higher dropout rate. The results for both-universities-dataset show scores in an intermediate value between U Talca and UAI, this could be expected as it was trained using data from both universities. Universidad de Talca All Variables showed a higher score in the logistic regression and neural network suggesting that the addition of the non-shared variables improved the performance, at least when considering these models. However, these differences showed no significant difference between the scores in comparison to the U Talca dataset.

TABLE 6.12:  $F_1$  score + class, for each dataset.

| Model               | both            | UAI             | U Talca         | U Talca All Vars |
|---------------------|-----------------|-----------------|-----------------|------------------|
| Random model        | $0.27 \pm 0.02$ | $0.26 \pm 0.03$ | $0.31 \pm 0.04$ | $0.29 \pm 0.04$  |
| KNN                 | $0.35 \pm 0.03$ | $0.30 \pm 0.05$ | $0.42 \pm 0.05$ | -                |
| SVM                 | $0.36 \pm 0.02$ | $0.31 \pm 0.05$ | $0.42 \pm 0.03$ | -                |
| Decision tree       | $0.33 \pm 0.03$ | $0.28 \pm 0.03$ | $0.41 \pm 0.05$ | $0.41 \pm 0.05$  |
| Random forest       | $0.35 \pm 0.03$ | $0.30 \pm 0.06$ | $0.41 \pm 0.05$ | $0.40 \pm 0.04$  |
| Gradient boosting   | $0.37 \pm 0.03$ | $0.31 \pm 0.04$ | $0.41 \pm 0.05$ | $0.40 \pm 0.04$  |
| Naive Bayes         | $0.34 \pm 0.02$ | $0.29 \pm 0.04$ | $0.42 \pm 0.03$ | -                |
| Logistic regression | $0.35 \pm 0.03$ | $0.30 \pm 0.05$ | $0.41 \pm 0.03$ | $0.43 \pm 0.04$  |
| Neural network      | $0.35 \pm 0.03$ | $0.28 \pm 0.02$ | $0.39 \pm 0.05$ | $0.42 \pm 0.04$  |

Table 6.13 shows the  $F_1$  score – class for all models and datasets. The scores are higher than in the positive class, which was expected since the negative class corresponds to the majority class (non-dropout students). Even though we balanced the data when training, the test data (and the real-world data) is still unbalanced, which may have an influence. Similarly to the  $F_1$  score + class, it is also difficult to select a single model as the best one, since random forest could be considered the best in the both and UAI datasets, but, surprisingly, KNN had better performance on Universidad de Talca (which include Universidad de Talca All variables). On the contrary, even though it could be difficult to discard a model, neural network had one of the lowest performances among all models. When comparing the performance by dataset, it seems that the both dataset have higher scores (unlike the previous measure, where it had an intermediate value). Universidad de Talca scores were similar when including non-shared variables, but random forest surprises with a lower average score (but no significant difference), which could be explained by the fact that the model selects variables in each tree. Then, the selection of these new variables instead of the most important variables, such as mat, could affect negatively the performance of the model.

TABLE 6.13:  $F_1$  score - class, for each dataset.

| Model               | both      | UAI       | U Talca   | U Talca All Vars |
|---------------------|-----------|-----------|-----------|------------------|
| Random model        | 0.63±0.02 | 0.64±0.01 | 0.63±0.04 | 0.61±0.03        |
| KNN                 | 0.73±0.02 | 0.72±0.02 | 0.76±0.02 | -                |
| SVM                 | 0.76±0.02 | 0.69±0.04 | 0.71±0.03 | -                |
| Decision tree       | 0.79±0.03 | 0.78±0.04 | 0.73±0.03 | 0.72±0.04        |
| Random forest       | 0.80±0.02 | 0.82±0.01 | 0.74±0.03 | 0.72±0.03        |
| Gradient boosting   | 0.80±0.01 | 0.73±0.02 | 0.73±0.04 | 0.73±0.03        |
| Naive Bayes         | 0.77±0.01 | 0.68±0.03 | 0.74±0.02 | -                |
| Logistic regression | 0.73±0.01 | 0.72±0.03 | 0.73±0.01 | 0.74±0.03        |
| Neural network      | 0.76±0.03 | 0.67±0.01 | 0.73±0.03 | 0.72±0.08        |

The next performance measure is accuracy, which is shown in Table 6.14. The results show that random forest could be considered the best in both and UAI datasets, but KNN had better performance on Universidad de Talca (which includes Universidad de Talca All variables), and again, some differences are not even significant. In contrast, again, the neural network could be selected as a model to discard, since it had the lowest score among the whole table. When comparing the performance by dataset, it seems similar to the  $F_1$  score – class behavior, with the both dataset having the higher scores with the same models. Random forest U Talca All Vars also showed a lower average score (but no significant difference) compared to Universidad de Talca.

TABLE 6.14: Accuracy, for each dataset.

| Model               | both      | UAI       | U Talca   | U Talca All Vars |
|---------------------|-----------|-----------|-----------|------------------|
| Random model        | 0.51±0.02 | 0.51±0.01 | 0.52±0.04 | 0.49±0.03        |
| KNN                 | 0.62±0.02 | 0.60±0.02 | 0.66±0.03 | -                |
| SVM                 | 0.65±0.02 | 0.57±0.05 | 0.61±0.03 | -                |
| Decision tree       | 0.68±0.03 | 0.66±0.05 | 0.63±0.03 | 0.63±0.05        |
| Random forest       | 0.69±0.02 | 0.71±0.02 | 0.64±0.03 | 0.62±0.04        |
| Gradient boosting   | 0.69±0.02 | 0.61±0.03 | 0.63±0.04 | 0.63±0.03        |
| Naive Bayes         | 0.66±0.01 | 0.56±0.04 | 0.64±0.03 | -                |
| Logistic regression | 0.62±0.02 | 0.60±0.03 | 0.63±0.02 | 0.64±0.03        |
| Neural network      | 0.66±0.03 | 0.55±0.10 | 0.63±0.08 | 0.63±0.07        |



The AUC, at Table 6.15, showed similar performance between all the models, which can be specially noticed in the U Talca dataset. Even without considering significant differences, no model seems to be better or worse than the others. When comparing the performance by dataset, as with the models, no dataset seems to behave better or worse than the others. Universidad de Talca scores were similar when including non-shared variables, with exception of the logistic regression and neural network, with an improvement that could be significant in the latter. As in the  $F_1$  score + class measure, the addition of new variables could have revealed new patterns found by that models.

TABLE 6.15: AUC, for each dataset.

| Model               | both      | UAI       | U Talca   | U Talca All Vars |
|---------------------|-----------|-----------|-----------|------------------|
| Random model        | 0.50±0.04 | 0.51±0.03 | 0.52±0.06 | 0.48±0.05        |
| KNN                 | 0.64±0.03 | 0.61±0.04 | 0.66±0.03 | -                |
| SVM                 | 0.64±0.02 | 0.62±0.06 | 0.67±0.02 | -                |
| Decision tree       | 0.63±0.03 | 0.61±0.05 | 0.66±0.03 | 0.65±0.03        |
| Random forest       | 0.65±0.03 | 0.62±0.05 | 0.66±0.03 | 0.66±0.03        |
| Gradient boosting   | 0.66±0.03 | 0.63±0.03 | 0.65±0.04 | 0.66±0.02        |
| Naive Bayes         | 0.65±0.03 | 0.63±0.04 | 0.66±0.02 | -                |
| Logistic regression | 0.65±0.02 | 0.62±0.05 | 0.66±0.03 | 0.68±0.03        |
| Neural network      | 0.65±0.03 | 0.63±0.03 | 0.66±0.02 | 0.69±0.03        |

As a summary, all results show similar performance among most models in each dataset. If it were necessary to select one model for the implementation of a dropout prevention system, we would select gradient boosting decision tree because we prioritize the scores with the  $F_1$  score + class measure, since the data was highly unbalanced. Recall that the  $F_1$  score + class would focus on classifying correctly the students who drop out (keeping a balance with the other classification), without achieving a high score when labeling all students as if they do not drop out (which are most of the students).

The influence of individual variables was examined through the interpretable models. Recall, the pattern to read the importance of the variable in Table 6.16 is “both,UAI,UTalca,Utalca All vars” and the values Y or N implies the use of that variable for the model. Note that in the last dataset there were only considered the models that had a difference with the Universidad de Talca dataset.

Given all models, the most important variable is mat. This variable was considered by practically all models except by a single case (UAI-Logistic regression). In this case, the variable pps could have included part of the information of mat, since it had a strong negative  $\beta$  value, and probably the addition of region affected in some way (since this is the only model where it is used). The second most important variables are pps and lang, which are shared by most models, but not for all datasets. Naive Bayes did not consider these variables (except for pps in the both dataset, where the unification of datasets may be the reason for its use), and they were mostly considered in the both and UAI datasets. This could be explained since the conditional distribution of the classes is sufficiently similar to not be considered by the model, or simply since they were not selected in the tuning process. Ranking was considered in some datasets in all of the models with exception of logistic regression, which did not consider this variable in any dataset. It was probably not used in some models due to collinearity with variables like pps or nem. Optional showed something similar, some datasets considered this variable important, but the decision tree did not consider it in any dataset. Recall that optional was the unification of 2 optional tests (history or science), which may have affected its use for prediction. Note that gradient boosting considered all of the previous variables as important in every dataset. Since this model predicts using the errors, it would be expected that using numerical variables, which have many possible values, would be preferable, instead of using one-hot encoded



categorical variables. The variable *nem* was considered important in few models, which was expected, since it was discarded from some models due to collinearity problems, specifically with the variable ranking.

The discrete variables were considered important only in a few models and datasets: preference and admission in UTalca-Logistic regression, where a lower value in preference (higher preference for the university) and entering using PSU decreases dropout, and region in UAI-Logistic regression, where being from region Metropolitana or Valparaíso decreases dropout. The variables degree and fam income were non-shared variables included in U de Talca All vars, however, the fact that they were considered important in one or two models may imply that some patterns could not be found using only the shared variables. Note that in U Talca the variables preference and admission seem to be replaced by the variable degree. Recall that degree was one-hot encoded, and the model considered important the degrees civil mecánica (which decreases dropout), civil en obras civiles (also decreases dropout), and bioinformática (which increases dropout). Finally, when comparing the universities, it seems that the feature importance in UAI defined the importance in the both-universities-dataset in most of the cases, which may be because the dataset of this university is bigger than the U de Talca dataset.

From a qualitative point of view, the importance of some variables in the first-year dropout could be explained. The variable *mat* could be the most important variable due to courses like calculus or physics, which uses several mathematical logic. As a fact, UAI realizes many efforts during the first undergraduate year to improve the mathematical knowledge of their students, for example, study groups before all the tests. These courses also used to have high failure rates. So, it will not be surprising that low mathematical scores will be positively related to failures in mathematical courses. The importance of *pps* could be explained by its relation with all other variables. Recall that *pps* is a weighted average of all other tests, being math one of the most important tests. Then, a high math score should imply a high score in *pps* too.

The variable *lang* has a different behavior than previous variables. In interpretative models, a higher score in *lang* increases the probability of dropout. At UAI, this could be explained by the low quantity of language-related courses and their low failure rates. From the 12 first-year subjects, only 3 of them evaluate reading and/or writing skills, with even a lower number of credits, than other courses. In contrast, all other courses, including mathematics, physical education, physic, and programming, language skills are not evaluated.

The variable ranking is a score that compares the student with other students from their school. Therefore, it seems reasonable that good students in school continue to be good students in university, hence its importance. It is common to observe that good students became friends and start generating study habits since the first semester. For example, using their free time between classes to study or solve some homework.

On the discrete variables, it is striking that the variable region was considered important only for UAI. In this case, students from the same region that the university, which usually lives with their parents, have a lower probability of dropout. This could be expected given the implications of changing residence at a shortage. First-year students that live alone have more freedom and responsibilities. In several cases, this freedom could imply more recreational parties and depression (due to loneliness), affecting negatively their performance during the first year.

TABLE 6.16: Feature importance, for each model and dataset. The pattern of each cell represents the datasets “both,UAI,UTalca,Utalca All vars”

| Var        | Decision tree | Random forest | Gradient boosting | Naive Bayes | Logistic reg. |
|------------|---------------|---------------|-------------------|-------------|---------------|
| mat        | Y,Y,Y,Y       | Y,Y,Y,Y       | Y,Y,Y,Y           | Y,Y,Y,-     | Y,N,Y,Y       |
| pps        | Y,Y,N,N       | Y,Y,N,N       | Y,Y,Y,Y           | Y,N,N,-     | Y,Y,Y,N       |
| lang       | Y,Y,Y,N       | Y,Y,N,N       | Y,Y,Y,Y           | N,N,N,-     | Y,Y,Y,N       |
| ranking    | N,N,Y,Y       | Y,Y,N,N       | Y,Y,Y,Y           | Y,Y,N,-     | N,N,N,N       |
| optional   | N,N,N,N       | Y,Y,N,N       | Y,Y,Y,Y           | Y,Y,N,-     | Y,N,N,N       |
| nem        | N,N,N,N       | N,N,N,N       | N,N,N,N           | Y,N,Y,-     | N,N,Y,Y       |
| degree     | -,-,-,N       | -,-,-,N       | -,-,-,Y           | -,-,-,-     | -,-,-,Y       |
| preference | N,N,N,N       | N,N,N,N       | N,N,N,N           | N,N,N,-     | N,N,Y,N       |
| admission  | N,N,N,N       | N,N,N,N       | N,N,N,N           | N,N,N,-     | N,N,Y,N       |
| region     | N,N,N,N       | N,N,N,N       | N,N,N,N           | N,N,N,-     | N,Y,N,N       |
| fam income | -,-,-,N       | -,-,-,N       | -,-,-,Y           | -,-,-,-     | -,-,-,N       |

## Chapter 7

# Conclusion

This work compared the performance and learned patterns from machine learning models over two universities when predicting student dropout of first-year engineering students. Four different datasets were compared: “both-universities-dataset” (students from both of the universities and shared variables), “Universidad Adolfo Ibáñez” (students from this university and all variables, which are the same as the shared variables), “Universidad de Talca” (students from this university and the shared variables), and “Universidad de Talca All variables” (the same than Universidad de Talca, but included non-shared variables).

From the numerical perspective, the results show similar performance among most models in each dataset, separately. If it were necessary to select one model for the implementation of a dropout prevention system, we would prioritize the scores with the  $F_1$  score + class measure, since the data was highly unbalanced. Considering this, the best option would be gradient boosting decision tree, since it showed the higher average score in the both and UAI datasets, with good scores in the Universidad de Talca and Universidad de Talca All variables datasets. Following that priority, it would be reasonable to discard the decision tree, for its lower average score when using that measure.

The interpretive models (decision tree, random forest, gradient boosting, naive Bayes, and logistic regression) showed that the most important variable is mat (mathematical PSU test score), since this variable was considered in almost every model and datasets. In all the cases, a higher score of this variable decreased the probability of dropout. The importance of this variable makes sense since many of the efforts done inside the university during the first year are focused on courses like calculus or physics, which are mathematical courses (e.g. study groups organized by the university or student organizations), and these courses also use to have high failure rates. Other variables, like pps and ranking, were also considered by most models, where a higher score also decreased the probability of dropout. The variable lang was considered by some models too, but surprisingly a higher score increased the probability of dropout, which could be explained by the fact that we were analyzing engineering students, and reading and writing skills are barely evaluated during first year. On the opposite, most categorical variables were not considered important by most of the models. The few exceptions are preference and admission in Universidad de Talca dataset, region in Universidad Adolfo Ibáñez, and family income and degree as non-shared variables in Universidad de Talca, where the last variable seems to replace the information of preference and admission, which shows a limitation of the unification of datasets. Most of these variables were selected in Universidad de Talca, suggesting differences in the universities.

Finally, when comparing among universities, the unification of datasets resulted in an intermediate performance (between the score of the 2 universities) in 2 of 4 measures, revealing that one university would become a limitation in the performance of the other when a general model is used. For that, it would be preferable to use a single model per university, instead of a general model. A single model would focus better on the patterns of each university, while a general model may lose information when trying to generalize them.

As future work, it would be important to collect more (and different) variables to include in the models. A better model could be generated using data related to the adaptation and social integration information of the students, which was used in older studies. Additionally, if required to predict dropout during the semester (and not only before entering), it would be useful to collect day-to-day information through learning content managements like Moodle, or the one available in the university of interest.

# Bibliography

Ministerio de educación, Chile

David Delen

- [1] D. Delen, "A comparative analysis of machine learning techniques for student retention management," *Decision Support Systems*, vol. 49, no. 4, p. 498–506, 2010.

W. H. Spady

- [2] W. Spady, "Dropouts from higher education: An interdisciplinary review and synthesis," *Interchange*, vol. 1, pp. 64–85, 1970.

V. Tinto

- [3] V. Tinto, "Dropout from higher education: A theoretical synthesis of recent research," *Review of Educational Research*, vol. 45, no. 1, pp. 89–125, 1975.

L. C. J. B. P. A. M. Gonzalez

- [4] L. Cabrera, J. Bethencourt, P. Álvarez, and M. González, "El problema del abandono de los estudios universitarios. [the dropout problem in university study]," *Revista Electrónica de Investigación y Evaluación Educativa*, vol. 12, 2006.

- [5] M. Broc, "Voluntad para estudiar, regulación del esfuerzo, gestión eficaz del tiempo y rendimiento académico en alumnos universitarios," *Revista de Investigación Educativa*, vol. 29, pp. 171–185, 01 2011.

- [6] J. Bean, "Student attrition, intentions, and confidence: Interaction effects in a path model," *Research in Higher Education*, vol. 17, pp. 291–320, 1981.

- [7] E. Pascarella and P. Terenzini, *How college affects students: Findings and insights from twenty years of research*. ERIC, 1991.

- [8] L. Bejarano, S. Arango, K. Johana, H. Durán, and C. Ortiz, "Caso de estudio: caracterización de la deserción estudiantil en la fundación universitaria los libertadores 2014-1 – 2016-1," *Revista Tesis psicológica*, vol. 12, no. 2, pp. 138–161, 2017.

- [9] E. Sinchi and G. Ceballos, "Acceso y deserción en las universidades. alternativas de financiamiento," *Alteridad*, vol. 13, pp. 274–287, 06 2018.

- [10] I. Quintero, "Análisis de las causas de deserción universitaria," 2016.

- [11] B. Minaei-Bidgoli, D. Kashy, G. Kortemeyer, and W. Punch, "Predicting student performance: an application of data mining methods with an educational web-based system," vol. 1, pp. T2A–13, 12 2003.

- [12] A. Bernardo, R. Cerezo, J. Núñez, E. Tuero, and M. Esteban, "Prediction of university drop-out: Explanatory variables and preventive measures," *Revista Fuentes*, vol. 16, pp. 63–84, 2015.

- [13] T. Larroucau, "Estudio de los factores determinantes de la deserción en el sistema universitario chileno," *Revista Estudios de Políticas Públicas*, vol. 1, no. 1, pp. 1–23, 2015.

- [14] H. Kuna, R. Garcia-Martinez, and F. Villatoro, "Pattern discovery in university students desertion based on data mining," *Advances and Applications in Statistical Sciences*, vol. 2, 01 2010.

- [15] A. Garzón and J. Gil, "El papel de la procrastinación académica como factor de la deserción universitaria," *Revista Complutense de Educación*, vol. 28, no. 1, pp. 307–324, 2016.

- [16] P. Jia and T. Maloney, "Using predictive modelling to identify students at risk of poor university outcomes," *Higher Education*, vol. 70, no. 1, pp. 127–149, 2015.
- [17] R. Martelo, D. A. and P. Martelo, "Análisis multivariado aplicado a determinar factores clave de la deserción universitaria," *Revista Espacios*, vol. 39, no. 10, p. 13, 2018.
- [18] P. Giovagnoli, "Determinants in university desertion and graduation: An application using duration models," 2005.
- [19] C. Vallejos and M. Steel, "Bayesian survival modelling of university outcomes," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 180, no. 2, pp. 613–631, 2017.
- [20] J. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.
- [21] S. Kumar, B. Bharadwaj, and S. Pal, "Mining education data to predict student's retention: A comparative study," *International Journal of Computer Science and Information Security*, vol. 10, pp. 113–117, 03 2012.
- [22] D. Heredia, Y. Amaya, and E. Barrientos, "Student dropout predictive model using data mining techniques," *IEEE Latin America Transactions*, vol. 13, no. 9, pp. 3127–3134, 2015.
- [23] P. Ramírez-Correa and E. Grandón, "Predicción de la deserción académica en una universidad pública chilena a través de la clasificación basada en Árboles de decisión con parámetros optimizados," *Formación universitaria*, vol. 11, pp. 3–10, 06 2018.
- [24] D. R. Cox, "The regression analysis of binary sequences," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 20, no. 2, pp. 215–232, 1958.
- [25] A. Cabrera, "Logistic Regression Analysis in Higher Education: An Applied Perspective," vol. 10, pp. 225–256. 01 1994.
- [26] V. Santelices, X. Catalán, C. Horn, and D. Kruger, "Determinantes de deserción en la educación superior chilena, con Énfasis en efecto de becas y créditos," 2013.
- [27] A. Matheu, C. Ruff, M. Ruiz, L. Benites, and G. Morong, "Modelo de predicción de la deserción estudiantil de primer año en la universidad bernardo o'higgins," *Educação e Pesquisa*, vol. 44, 06 2018.
- [28] P. Langley, W. Iba, and K. Thompson, "An analysis of bayesian classifiers," in *In proceedings of the Tenth national conference on Artificial Intelligence*, pp. 223–228, MIT Press, 1992.
- [29] B. Kumar and S. Pal, "Data mining : A prediction of performer or underperformer using classification," (*IJCSIT*) *International Journal of Computer Science and Information Technologies*, vol. 9, no. 4, 2011.
- [30] V. Hegde and P. Prageeth, "Higher education student dropout prediction and analysis through educational data mining," in *2018 2nd International Conference on Inventive Systems and Control (ICISC)*, pp. 694–699, 2018.
- [31] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.
- [32] T. Tanner and H. Toivonen, "Predicting and preventing student failure – using the k-nearest neighbour method to predict student performance in an online course environment," *IJLT*, vol. 5, pp. 356–377, 01 2010.
- [33] M. Mardolkar and N. Kumaran, "Forecasting and avoiding student dropout using the k-nearest neighbor approach," *SN Computer Science*, vol. 1, 03 2020.

- [34] G. Zhang, "Neural networks for classification: a survey," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 30, no. 4, pp. 451–462, 2000.
- [35] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, pp. 206–215, 05 2019.
- [36] A. Siri, "Predicting students' dropout at university using artificial neural networks," *Italian Journal of Sociology of Education*, vol. 7, no. 2, pp. 225–247, 2015.
- [37] M. Alban and D. Mauricio, "Neural networks to predict dropout at the universities," *International Journal of Machine Learning and Computing*, vol. 9, pp. 149–153, 04 2019.
- [38] E. Fischer, "Modelo para la automatización del proceso de determinación de riesgo de deserción en alumnos universitarios," Master's thesis, Universidad de Chile - Facultad de Ciencias Físicas y Matemáticas, 2012.
- [39] K. Eckert and R. Suenaga, "Análisis de deserción-permanencia de estudiantes universitarios utilizando técnica de clasificación en minería de datos," *Formación universitaria*, vol. 8, pp. 03–12, 12 2014.
- [40] M. Miranda and J. Guzmán, "Análisis de la deserción de estudiantes universitarios usando técnicas de minería de datos," *Formación universitaria*, vol. 10, pp. 61–68, 01 2017.
- [41] A. Vilorio, J. Garcia, C. Vargas-Mercado, H. Hernández-Palma, N. Orellano, and M. Arrozola, "Integration of data technology for analyzing university dropout," *Procedia Computer Science*, vol. 155, pp. 569–574, 01 2019.
- [42] L. Kemper, G. Vorhoff, and B. Wigger, "Predicting student dropout: A machine learning approach," *European Journal of Higher Education*, vol. 10, pp. 1–20, 01 2020.
- [43] C. Cortes and V. Vapnik, "Support-vector networks," in *Machine Learning*, pp. 273–297, 1995.
- [44] T. Ho, "Random decision forests," in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, vol. 1, p. 278, IEEE Computer Society, 1995.
- [45] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

## Appendix A

# Machine learning models

This section describes the main machine learning models used in the literature.

### A.1 K-Nearest Neighbors

Given a set of data points  $\mathcal{D} = \{\mathbf{X}_1, \dots, \mathbf{X}_{N_n}\}$ , the  $K$ -nearest neighbors algorithm classifies a data point based on the vote of its  $K$  closest “neighbors”, based on the idea that closer data points behave similarly. To classify a new data point  $\mathbf{X}_p$ , the  $K$  nearest neighbors of  $\mathbf{X}_p$  from  $\mathcal{D}$  are calculated using a function distance, for example the Euclidean distance. Then, the final classification is made based on the class labels of the majority of these  $K$  neighbors. The value of  $K$  is determined by the user, and significantly affects the performance of the algorithm. A lower  $K$  creates small neighborhoods, making it difficult to find general patterns, while a larger value creates large neighborhoods, where small patterns could be lost.

### A.2 Support vector machine

Originally used for two-groups classification problems ([43]), this method uses a hyperplane to separate between classes. In the case of an optimal classification, the classes are perfectly separated by the hyperplane. The algorithm searches for the hyperplane that maximizes the margin between the classes, classifying the data points according to their position with respect to the defined hyperplane. In the case of classes that are not linearly separable, a kernel is used to increase the dimension of the data points, finding a hyperplane in this new dimension. The optimization problem for training the algorithm can be summarized as:

$$\begin{aligned} & \text{minimize} && \frac{1}{2}w^\top w + C \sum_{i=1}^{N_n} \xi_i \\ & \text{subject to} && y_i [w^\top \phi(x_i) + b] \geq 1 - \xi_i \quad \forall i \in \{1, \dots, N_n\} \end{aligned} \quad (\text{A.1})$$

Where  $\xi_i > 0$  for  $i = 1, 2, \dots, M$  are the error margin associated with the classification cost  $C$ ,  $y_i$  are the resulting classification for the training set, and  $\phi(x_i)$  is the kernel function transforming the data to the new  $m$  dimensional space. To solve this, Lagrange multipliers are normally used, as it is an optimization problem that is meant to find the maximum margin between the different classes.

### A.3 Decision Trees

The decision trees are structures composed of nodes, leaves, and branches, where each node corresponds to a decision on an attribute to resolve how to classify a determined object, and leaves determine the final output ([20]). For example, in this thesis, a node could ask about a previous degree, and a leaf determines if a student will dropout. The output is easily visualized through a flowchart



in form of a “tree” (Figure A.1), where each branch represents a possible path of these decisions. When a data point is inserted into the tree, it travels through a path until a leaf is reached. Each leaf node represents the output probability for the input point of belonging to a determined class. The training process of a tree determines the best attributes for each node, generating the tree structure. Specifically, the training process searches for the attribute that minimize/maximize a scoring function based on the final classification (for example the gini index or entropy). This process is repeated until all attributes are selected, all the train points are classified, or the score function can not longer be improved. Once the training process is ready, a post pruning process is applied to reduce the size of the tree avoiding overfitting (memorizing the pattern of the training points instead of learning the real pattern of the data).

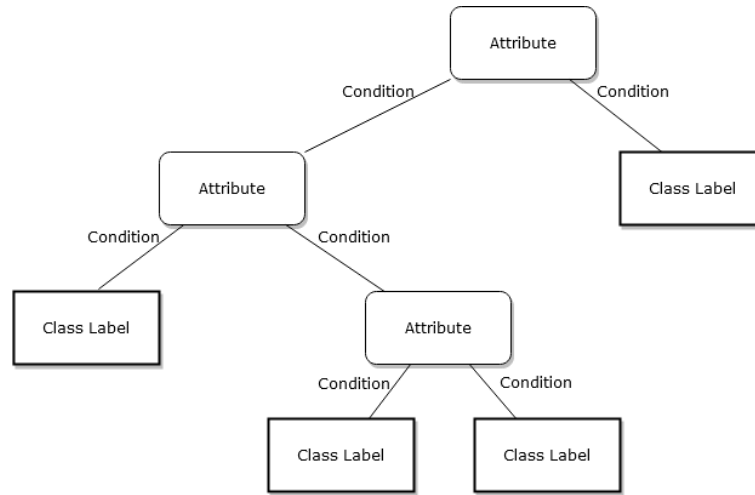


FIGURE A.1: Decision tree structure.

## A.4 Random forest

Following the principles of stochastic modeling, random forest is a method that constructs tree-based classifiers whose capacity can be arbitrarily expanded for increases in accuracy ([44]). It builds multiple decision trees (see A.3), each one using random samples of the variables. The different samples generalize their classification in complementary ways. The class label of a data point is determined using the mode of the classifications of these individual trees.

## A.5 Gradient boosting decision tree

A general gradient descent boosting paradigm is developed for additive expansions based on any fitting criterion ([45]). When used with decision trees, it uses regression trees to minimize the error of the prediction. A first tree predicts the probability of a data point to belong to a class, the next tree models the error of the first tree, minimizing it and calculating a new error, which is the new input for a new error-modeling tree. This boosting is supposed to improve the performance, where the final model is the sum of the output of each tree.

## A.6 Naive Bayes

The naive Bayes is a probabilistic model, based on the Bayes theorem and probability conditional independence ([28]). To estimate a value proportional to the probability of a class, it assumes independence among the set of variables  $\mathbf{X} = \{\mathbf{X}_1, \dots, \mathbf{X}_{N_m}\}$  conditioned on variable  $Y$ , where  $N_m$  is the

number of variables and  $Y$  is the class variable. However, given the difficulty of finding the conditional probability for each class label, the model assumes independence between the variables given the class, simplifying the expression, as seen in A.2.

$$P(Y|\mathbf{X}) = \frac{P(\mathbf{X}|Y)P(Y)}{P(\mathbf{X})} \propto \prod_{i=1}^{N_m} P(\mathbf{X}_i|Y)P(Y) \quad (\text{A.2})$$

With the learned parameters, the model calculates a value proportional to the probability of belonging to each class. To determine the class of a point, we simply select the class with the higher value.

## A.7 Logistic Regression

The logistic regression algorithm, first used by [24] as a probability model, is a non-linear regression that uses a sigmoid function to predict the probability of success for a random variable  $Y$  based on a set of random variables  $\mathbf{X} = \{\mathbf{X}_1, \dots, \mathbf{X}_{N_m}\}$  (A.3). Given a set of data points  $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{N_n}\}$ , where  $\mathbf{x}_i = \{x_{i1}, \dots, x_{iN_m}\}$  represents the value of the  $i$ -th point and  $x_{ij}$  is the value of the  $j$ -th attribute, the model learn a vector of parameters  $\boldsymbol{\beta} = \{\beta_0, \dots, \beta_{N_d}\}$  where  $\beta_j$  is the weight given to the  $j$ -th attribute, and  $\beta_0$  is the offset parameter.

$$P(Y|\mathbf{X}) = \frac{\exp(\beta_0 + \beta_1 \mathbf{X}_1 + \dots + \beta_{N_m} \mathbf{X}_{N_m})}{1 + \exp(\beta_0 + \beta_1 \mathbf{X}_1 + \dots + \beta_{N_m} \mathbf{X}_{N_m})} \quad (\text{A.3})$$

Positive values of  $\beta_j$  in equation (A.3) indicate that the attribute increases the probability of belonging to the positive class, while negative values decrease the probability. A unit increase in  $\mathbf{X}_j$  implies that the odds (ratio of positive class outcome to negative class outcome) are multiplied by  $\exp(\beta_j)$ . Then, a high  $|\beta_j|$ , for normalized values, implies a higher importance of the variable, either by largely increasing or decreasing the probability. In case of a classification problem, a success is usually assigned if the probability is greater than 0.5, otherwise is considered a failure.

## A.8 Artificial Neuronal Networks

The Artificial Neuronal Networks tries to simulate the behavior of biological neurons where the network learns how to perform a specific task. It is composed of many layers (each one with one or more neurons), where an input is delivered to the network (first layer), and the final layer (output layer) generates an output to make the classification decision. The inside process (hidden layers), for simplicity, can be seen as a black box. The train process adapts the connection among neurons to minimize a determined measure of error, for example, the mean squared error. In a classification problem, the output is usually one neuron per class, with the probability of belonging to that class. In the specific binary class problem, one neuron can be used for the output layer.

## Appendix B

# Variable description

After the initial cleaning process, we describe all variables of the final dataset. For each variable, we describe its meaning, distribution, and distribution conditioned on the drop-out variable. For this purpose, we used a table or plots, showing in the left position the distribution of the variable, while to the right of it the conditioned information.

### B.1 Universidad Adolfo Ibáñez

1. **ID ALUMNO**: Unique ID to identify each student.
2. **FECHA NACIMIENTO**: A numerical value indicating the year of birth, ranging from 1900 to 2012. A first analysis detected 15 outliers: 13 students born in 1900, 1 student born in 1918, and 1 student born in 2012, which were eliminated. The plot of Figure B.1 omitted the extreme values to improve visualization. We can observe that most of the values are between 1990 and 2000, which was expected considering that we deleted the values before COHORTE 2012, and that the average age to enter university is 18 years old. There is no significant difference between the distribution and the conditional distribution of the variable. Hence, this variable does not contribute to university first-year dropout, and it was eliminated.

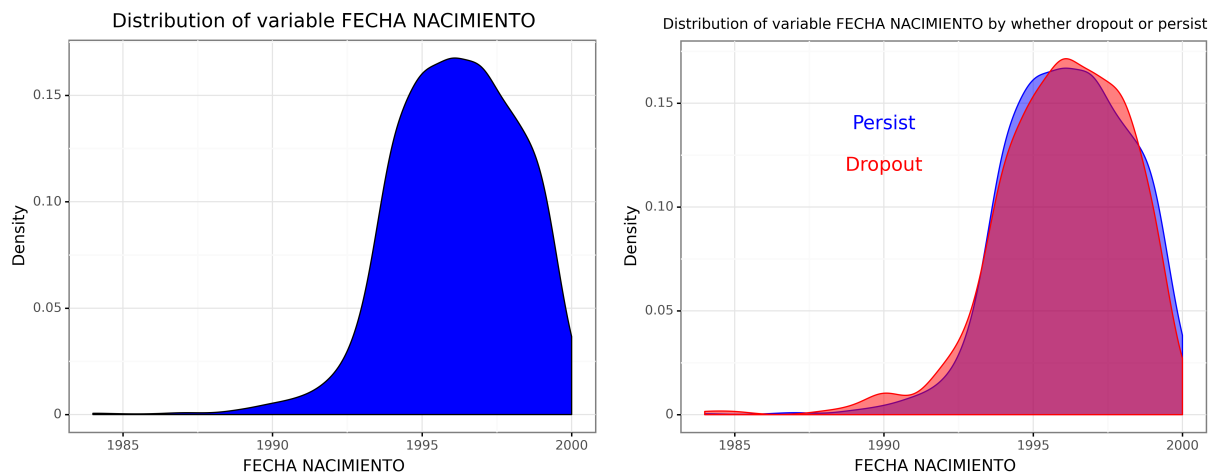


FIGURE B.1: (Conditional) Analysis distribution for FECHA NACIMIENTO: Similar behaviour between distribution and conditional distribution of the variable.

3. **GENERO**: Categorical variable that describes the gender of the students. It has 2 possible values, "M" or "F", indicating male or female respectively. We can see a slight increase in male dropout percent in relationship to total percent of male, which could be significant.

TABLE B.1: (Conditional) Analysis distribution for GENERO: slight increase in male dropout.

| Gender | Total frequency | Total percent | Dropout frequency | Dropout percent | Persist frequency | Persist percent |
|--------|-----------------|---------------|-------------------|-----------------|-------------------|-----------------|
| M      | 2893            | 77.15%        | 428               | 79.85%          | 2465              | 76.69%          |
| F      | 857             | 22.86%        | 108               | 20.15%          | 749               | 23.30%          |

4. **TIPO COLEGIO:** Categorical variable indicating the school type of the student. It has 3 possible values: Particular (private school), SUBVENCIONADO (private with state support), and Municipal (public school), with most of the values being PARTICULAR. We can see that students coming from public and private with state support schools have lower probabilities of dropout.

TABLE B.2: (Conditional) Analysis distribution for TIPO COLEGIO: lower dropout for Subvencionado and Municipal values.

| School dependency | Total frequency | Total percent | Dropout frequency | Dropout percent | Persist frequency | Persist percent |
|-------------------|-----------------|---------------|-------------------|-----------------|-------------------|-----------------|
| Particular        | 2856            | 76.16%        | 436               | 81.34%          | 2420              | 75.29%          |
| Subvencionado     | 538             | 14.35%        | 61                | 11.38%          | 477               | 14.84%          |
| Municipal         | 115             | 3.06%         | 13                | 2.42%           | 102               | 3.17%           |
| Nulls             | 241             | 6.43%         | 26                | 4.85%           | 215               | 6.69%           |

5. **TIPO COLEGIO 2:** Categorical variable indicating the school education type of the student. It has 5 possible values: Humanista Científica Diurno, Técnico Profesional Comercial, Humanista Científico Nocturno, Técnico Profesional Agrícola, and Técnico Profesional Comercial. Table B.3 shows only 9 students with a different value rather than Humanista Científica Diurno. Moreover, more than 20% of the values are null, and there are no values of students who have dropout. Given this information, we dropped this variable.

TABLE B.3: (Conditional) Analysis distribution for TIPO COLEGIO 2: No values for dropout.

| School type of education       | Total frequency | Total percent | Dropout frequency | Dropout percent | Persist frequency | Persist percent |
|--------------------------------|-----------------|---------------|-------------------|-----------------|-------------------|-----------------|
| Humanista Científica Diurno    | 2808            | 74.88%        | 0                 | 0 %             | 2808              | 87.37%          |
| Técnico Profesional Industrial | 4               | 0.11%         | 0                 | 0 %             | 4                 | 0.12%           |
| Humanista Científico Nocturno  | 3               | 0.08%         | 0                 | 0 %             | 3                 | 0.09%           |
| Técnico Profesional Agrícola   | 1               | 0.03%         | 0                 | 0 %             | 1                 | 0.03%           |
| Técnico Profesional Comercial  | 1               | 0.03%         | 0                 | 0 %             | 1                 | 0.03%           |
| Nulls                          | 933             | 24.88%        | 536               | 100 %           | 397               | 12.35%          |

6. **COHORTE:** Numerical variable indicating the admission year to the career. Values vary from 2013 to 2018 (after cleaning). Although was expected to be constant over the years, we can observe a slight increase in 2016 dropout in relationship to 2016 total values, which could be significant. This could be due to the higher amount of students who entered in 2016.

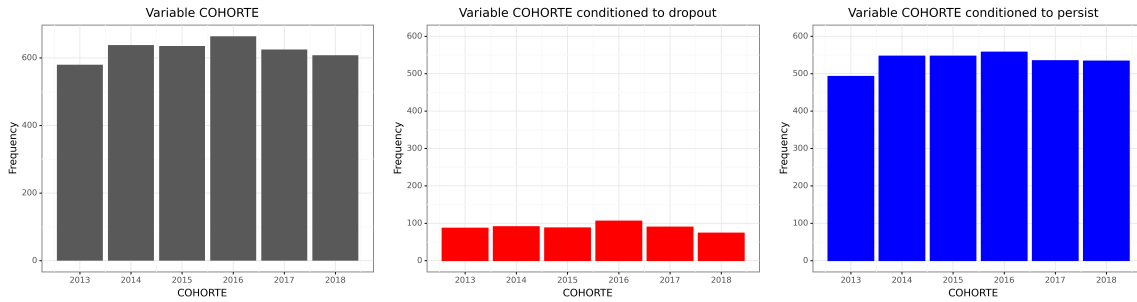


FIGURE B.2: (Conditional) Analysis distribution for COHORTE: We can observe a slight increase in 2016 dropout

7. **TIPO POSTULACION:** Categorical variable indicating the type of admission. It has 2 possible values: ADMISION REGULAR (regular admission), ADMISION ESPECIAL (special admission, such as sport scholarship). We can see that almost all of the values are ADMISION REGULAR, and a similar behavior among persist, dropout, and total percent.

TABLE B.4: (Conditional) Analysis distribution for TIPO POSTULACION: No significant difference between dropout and total percent.

| Type of admission | Total frequency | Total percent | Dropout frequency | Dropout percent | Persist frequency | Persist percent |
|-------------------|-----------------|---------------|-------------------|-----------------|-------------------|-----------------|
| Admisión regular  | 3457            | 92.19%        | 491               | 91.60%          | 2966              | 92.28%          |
| Admisión especial | 286             | 7.63%         | 43                | 8.02%           | 243               | 7.56%           |
| Nulls             | 7               | 0.19%         | 2                 | 0.37%           | 5                 | 0.16%           |

8. **Nota4Medio:** Numerical variable indicating the grade from the last year of high school. It ranges from 5.0 to 7.0. We can see that students who dropout have lower values than students who persist.

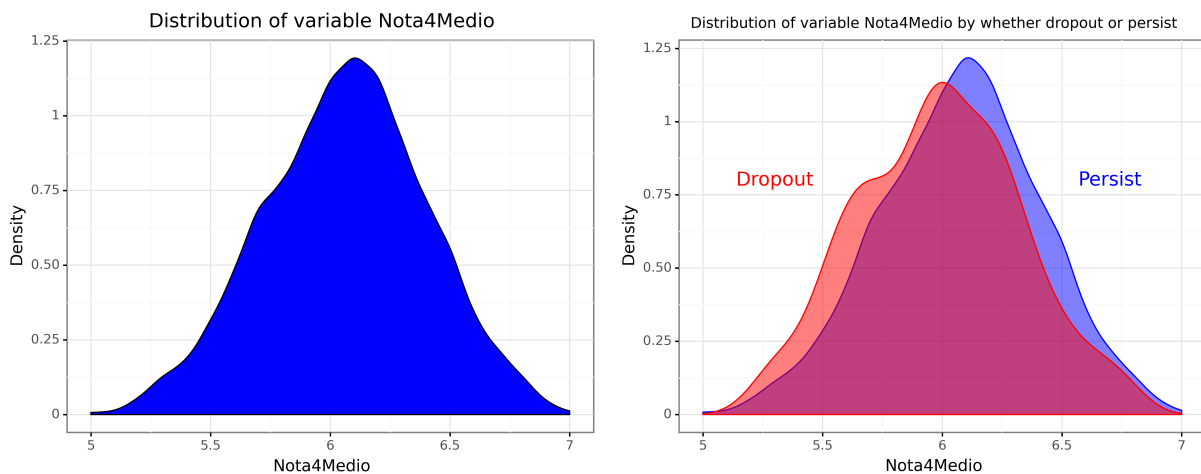


FIGURE B.3: (Conditional) Analysis distribution for Nota4Medio: lower values increases dropout.

9. **NEM:** Numerical variable that corresponds to the high school grades (average of the last 4 high school years, scaled to the range 0-850). It ranges from 414 to 822. The values seem like a normal distribution, with students who dropout having lower values than students who persist.

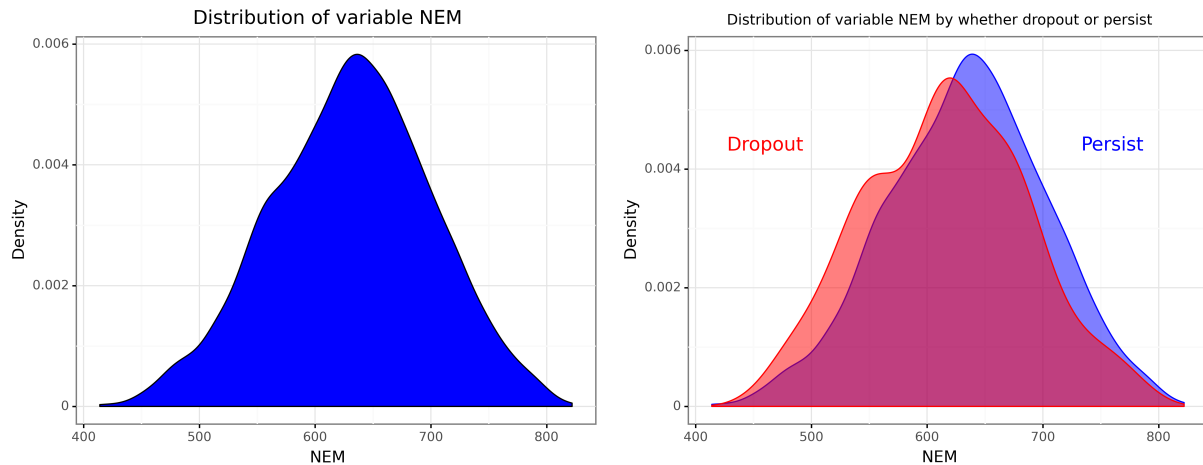


FIGURE B.4: (Conditional) Analysis distribution for Nem: lower values increases dropout.

10. **PSU MAT**: Numerical variable indicating the math test score. It ranges from 568 to 850. The values seem like a normal distribution, with students who dropout having considerably lower values than students who persist. This could lead to the left displacement of the curve in the dropout distribution.

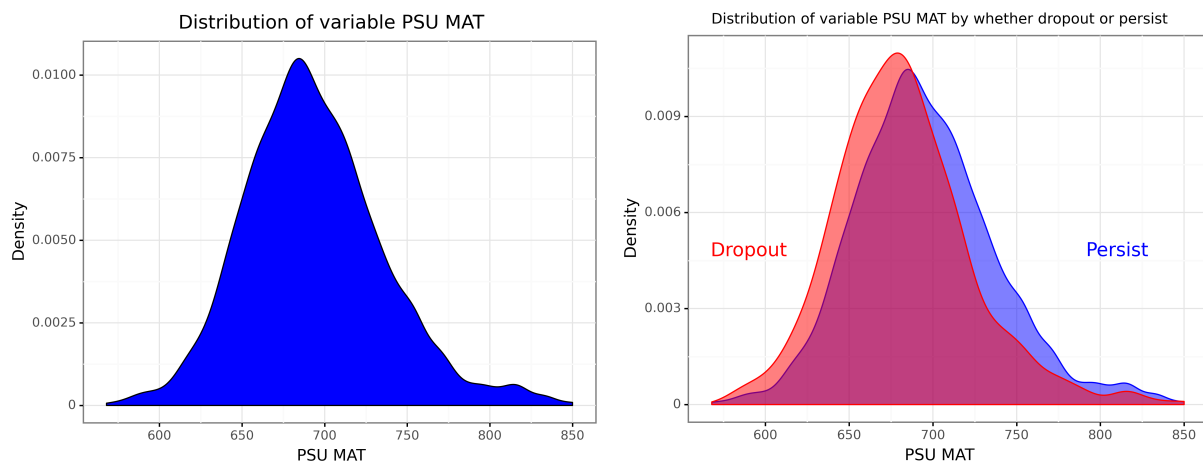


FIGURE B.5: (Conditional) Analysis distribution for PSU MAT: lower values increases dropout.

11. **PSU LENGUAJE**: Numerical variable indicating language test score. It ranges from 407 to 829. The values seem like a normal distribution. Although it was expected that students who dropout will have lower values than students who persist, this can not be appreciated. In fact, in the dropout distribution, we can see a heavier right tail.

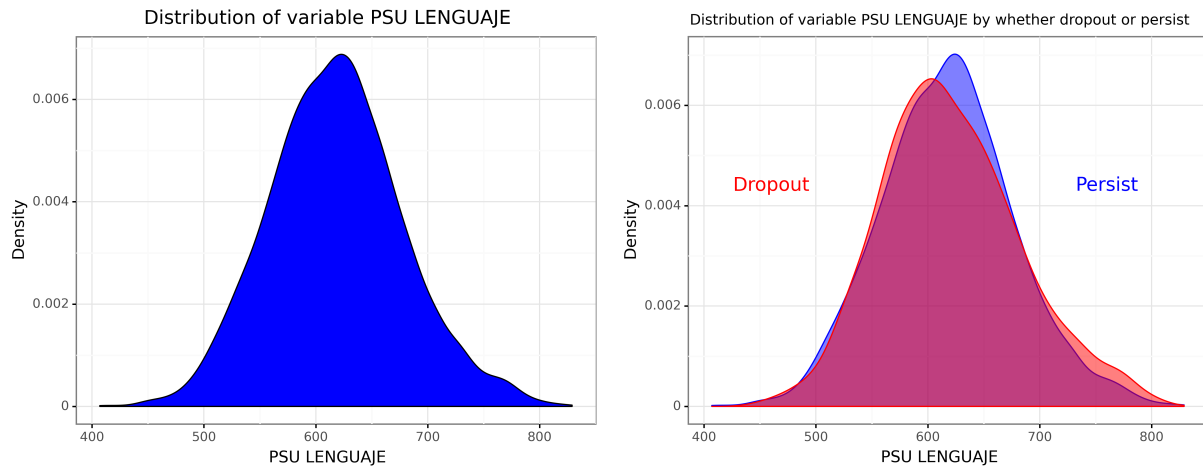


FIGURE B.6: (Conditional) Analysis distribution for PSU LENGUAJE: Higher dropout in higher values.

12. **PSU OPCIONAL**: Numerical variable indicating the maximum score between the history test and the science test. It ranges from 413 to 850. The values seem like a normal distribution, with students who dropout having similar values to students who persist.

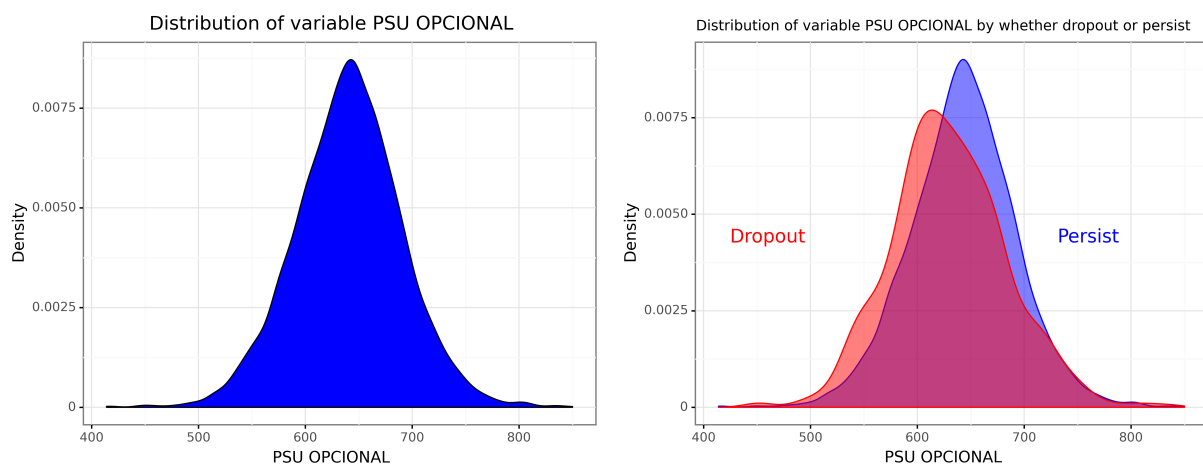


FIGURE B.7: (Conditional) Analysis distribution for PSU OPCIONAL: lower values increases dropout.

13. **PUNTAJE PONDERADO**: Numerical variable indicating the weighted student's university admission score. It ranges from 576 to 823, with only 8 values under the 600 points, probably exceptions of the minimum score requested to apply. We can see that students who dropout have lower values than students who persist.

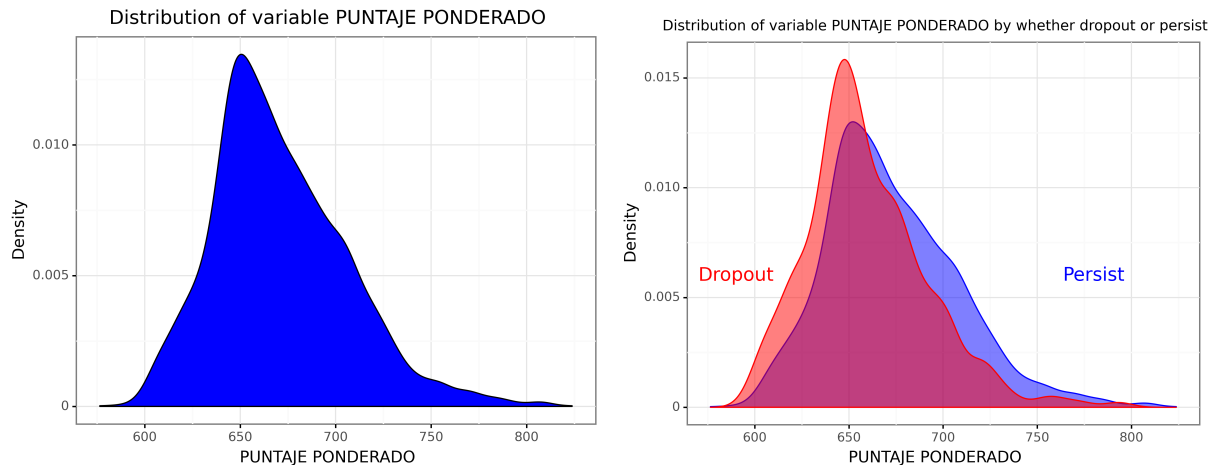


FIGURE B.8: (Conditional) Analysis distribution for PUNTAJE PONDERADO: lower values increases dropout.

14. **PUNTAJE RANKING:** Numerical variable indicating the student high school ranking, which is a selection factor that considers the academic performance of a student in relation to their educational context <sup>1</sup>. It has values from 414 to 850. The values seem like a normal distribution, except for a large number of 850 points (58 students). We can see that students who dropout have lower values than students who persist.

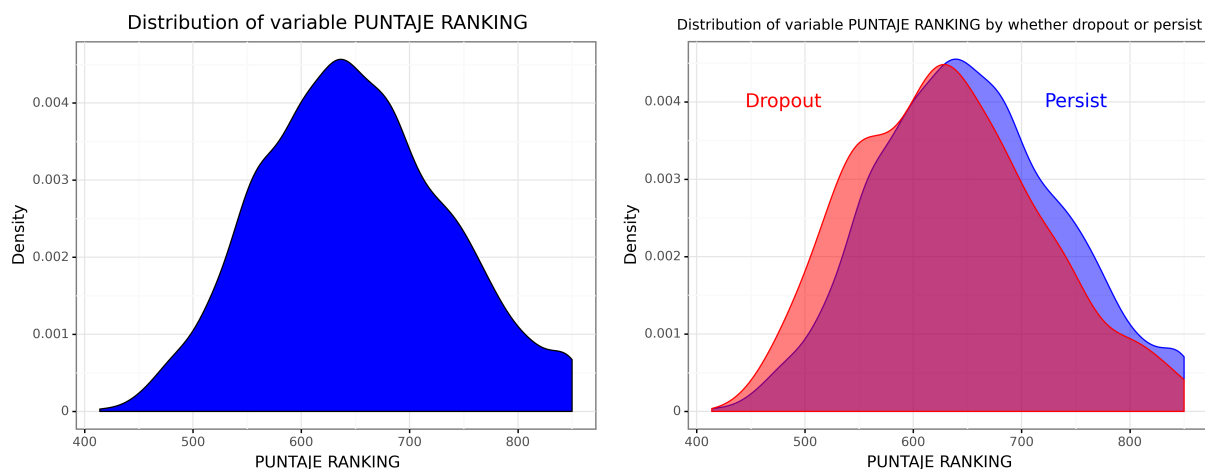


FIGURE B.9: (Conditional) Analysis distribution for PUNTAJE RANKING: lower values increases dropout.

15. **PREFERENCIA:** Numerical variable indicating the preference of the students for Universidad Adolfo Ibáñez in her/his application. It ranges from 1 to 10, where 1 means that Universidad Adolfo Ibáñez was the first option. We can see that the proportion of the values of preference 1 is lower in dropout.

<sup>1</sup><https://demre.cl/proceso-admision/factores-seleccion/puntaje-ranking>



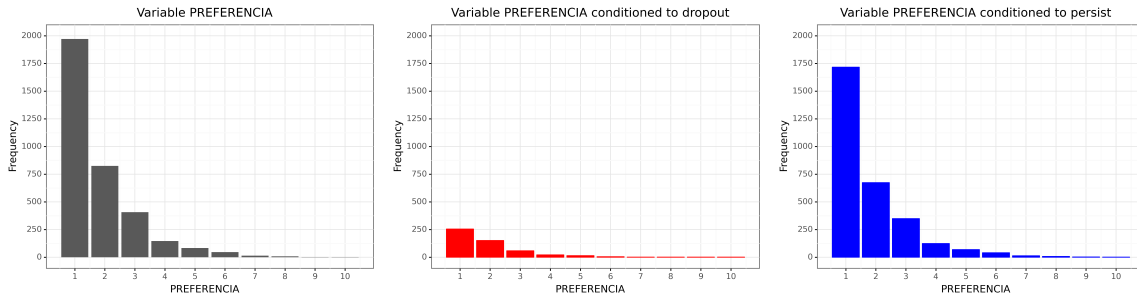


FIGURE B.10: (Conditional) Analysis distribution for PREFERENCIA: lower preference in students who dropout.

16. **COMUNA:** Categorical variable showing the commune of the student. It has 148 different communes, reduced to the 15 most frequent values and gathering the other values in the column Others. Although it can not be observed at simple sight, when comparing the proportions of the values, we observe a similar behavior between dropout and persist.

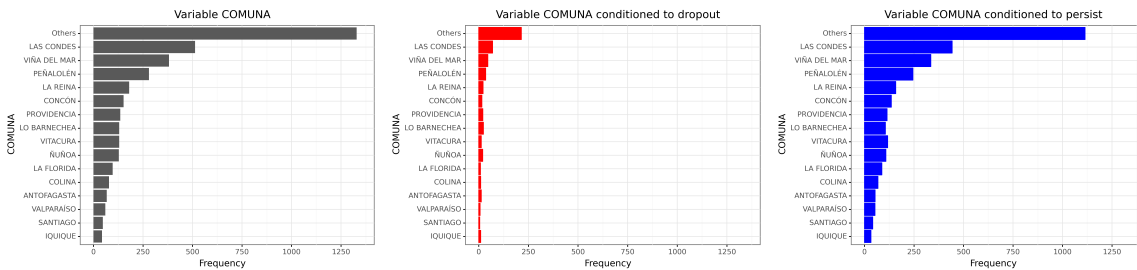


FIGURE B.11: (Conditional) Analysis distribution for COMUNA: Similar behaviour between dropout and persist.

17. **REGIÓN:** Categorical variable that corresponds to one of the 16 possible regions where the student comes from, grouping the northern and southern regions. Although it can not be observed at simple sight, when comparing the proportions of the values, we observe a slightly higher proportion in dropout in North and South regions, and similar behavior in RM and Valparaíso between dropout and persist.

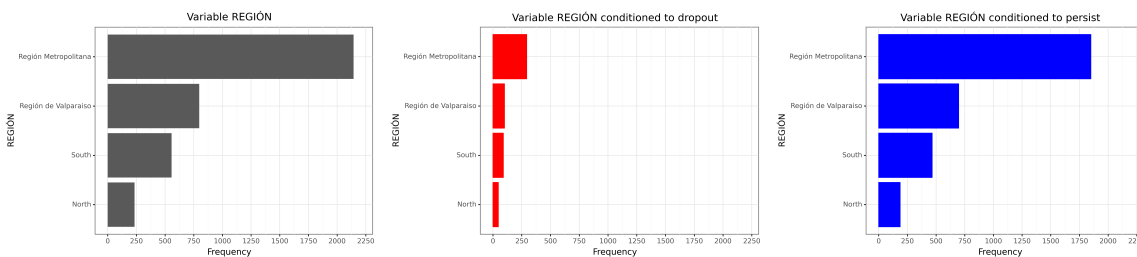


FIGURE B.12: (Conditional) Analysis distribution for REGIÓN: North and South regions with higher dropout.

18. **Semestres:** Categorical ordinal variable indicating the semester when the student drop out the university. Its possible values are 0, 0.5, 1, 2, 3, and 4, being 0 when the student has not dropout within the first two years, or the data is simply not available, and 0.5 when the student dropped

out before finishing the first semester. For the purposes of this analysis, it was binarized considering as 1 in case the student has dropped out in the first year (0.5 to 2) and 0 in other cases. The distribution of the variable for both cases is shown below.

TABLE B.5: Analysis distribution for Semestres (binarized).

| Semester | Frequency | Semester | Frequency |
|----------|-----------|----------|-----------|
| 0.0      | 3002      | 0        | 3204      |
| 0.5      | 77        | 1        | 535       |
| 1.0      | 157       |          |           |
| 2.0      | 301       |          |           |
| 3.0      | 119       |          |           |
| 4.0      | 83        |          |           |

## B.2 Universidad de Talca

1. Anho: Numerical value indicating the entry year of the student, ranging from 2003 to 2018, reduced to 2013 to 2017 after cleaning.
2. Rut: Numerical value indicating the unique ID of the students, conserved just for identification.
3. Fecha Nac: Numerical value indicating the year of birth, ranging from 1981 to 1999. There is no significant difference between the distribution and the conditional distribution of the variable. Hence, this variable does not contribute to first-year dropout and it was eliminated.

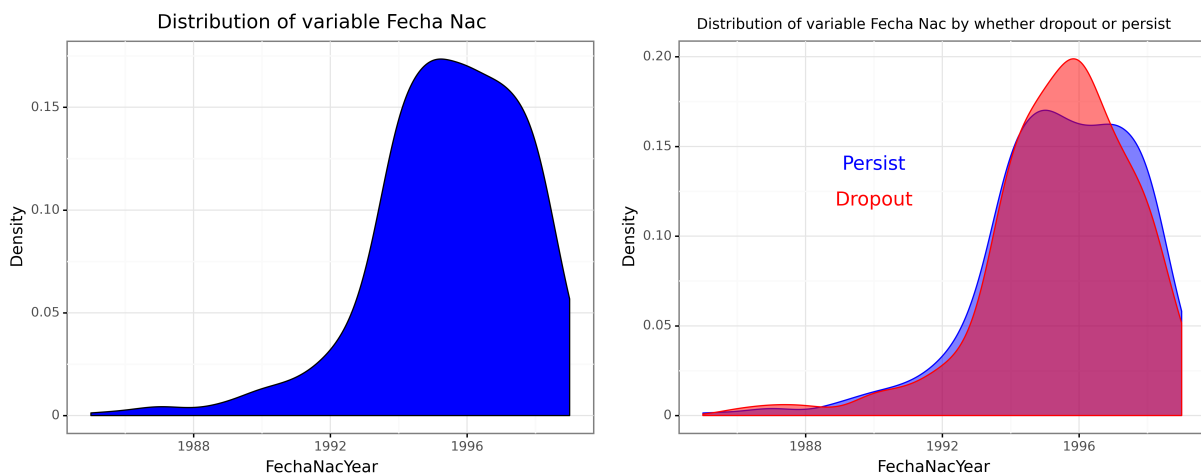


FIGURE B.13: (Conditional) Analysis distribution for Fecha Nac: Similar behaviour between distribution and conditional distribution of the variable

4. EntryAge: Numerical value indicating the age of the student when entered university, ranging from 18 to 37. It was created using the variables Anho and Fecha Nac. There is no significant difference between the distribution and the conditional distribution of the variable. Hence, this variable does not contribute to first-year dropout and it was eliminated.

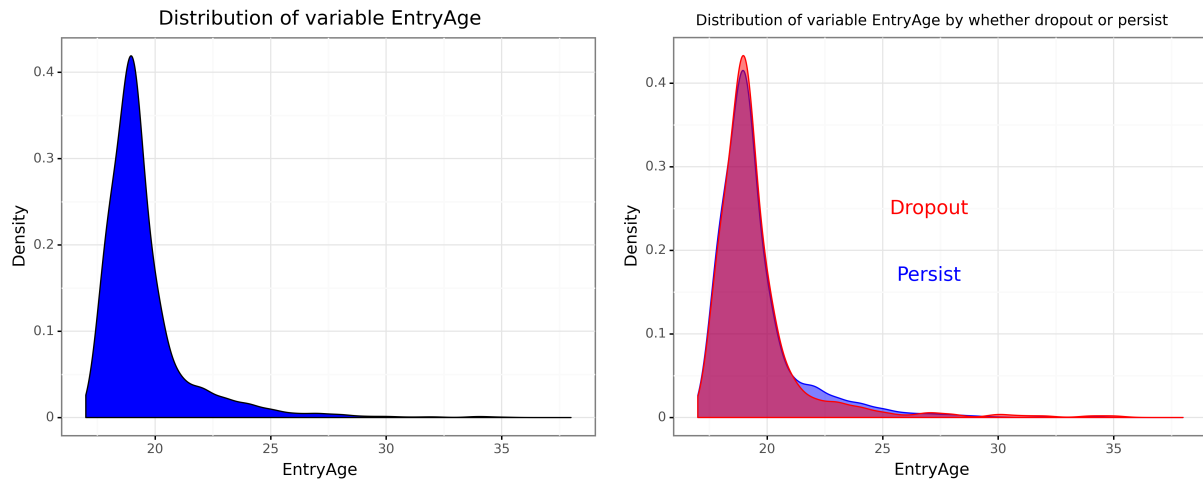


FIGURE B.14: (Conditional) Analysis distribution for EntryAge: Similar behaviour between distribution and conditional distribution of the variable

5. **Sexo:** Categorical variable that describes the gender of the students. It has 2 possible values, “M” or “F”, indicating male or female respectively. We can see a similar behavior among persist, dropout, and total percent. Although it is not observed that affects dropout, it was considered that, since this variable was maintained in UAI’s dataset, it would also be maintained here for future combined analyses.

TABLE B.6: Analysis distribution for Sexo: Similar behaviour among persist, dropout, and total percent.

| Gender | Total frequency | Total percent | Dropout frequency | Dropout percent | Persist frequency | Persist percent |
|--------|-----------------|---------------|-------------------|-----------------|-------------------|-----------------|
| M      | 1716            | 76.99%        | 366               | 76.25%          | 1350              | 77.19%          |
| F      | 513             | 23.01%        | 114               | 23.75%          | 399               | 22.81%          |

6. **Domicilio Ciudad:** Categorical variable showing the city of the student. It has 69 different cities, reduced to the 5 most frequent values, and gathering other values in the column Others. Although it can not be observed at simple sight, when comparing the proportions of the values, we observe a slightly higher proportion of dropout in LINARES, and a lower proportion of dropout in RANCAGUA.

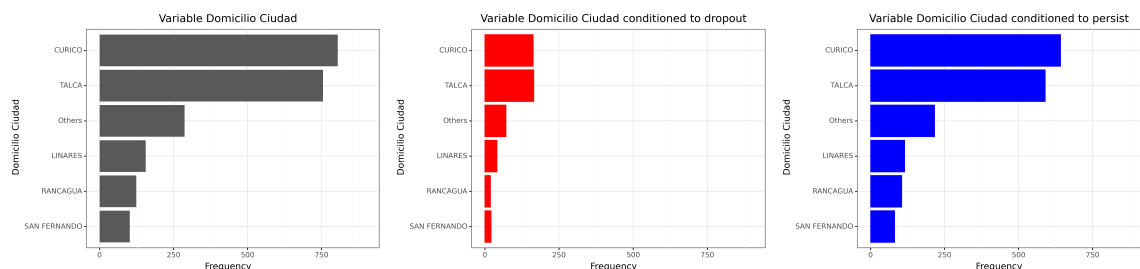


FIGURE B.15: (Conditional) Analysis distribution for Domicilio Ciudad: Higher dropout in LINARES, lower proportion of dropout in RANCAGUA

7. Domicilio Comuna: Categorical variable showing the commune of the student. It has 160 different cities, reduced to the 9 most frequent values and gathering other values in the column Others. We can see a similar behavior among total, dropout, and persist values.

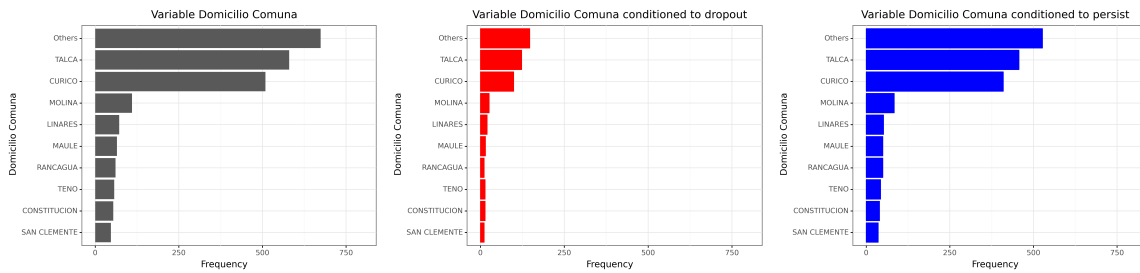


FIGURE B.16: (Conditional) Analysis distribution for Domicilio Comuna: Similar behavior among total, dropout, and persist values.

8. Domicilio Provincia: Categorical variable showing the province of the student. It has 43 different provinces, reduced to the 5 most frequent values, and gathering other values in the column Others. We can see a slightly higher dropout in Others.

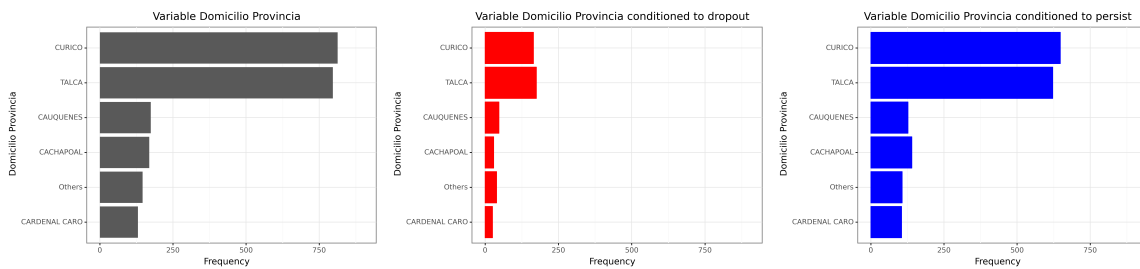


FIGURE B.17: (Conditional) Analysis distribution for Domicilio Provincia: Slightly higher dropout in Others.

9. Domicilio Region: Categorical variable that corresponds to one of the 16 possible regions where the student comes from, reduced to regions 6 and 7, gathering other regions in the column Others. We can see a slightly higher dropout in Others, and a lower dropout in 6.

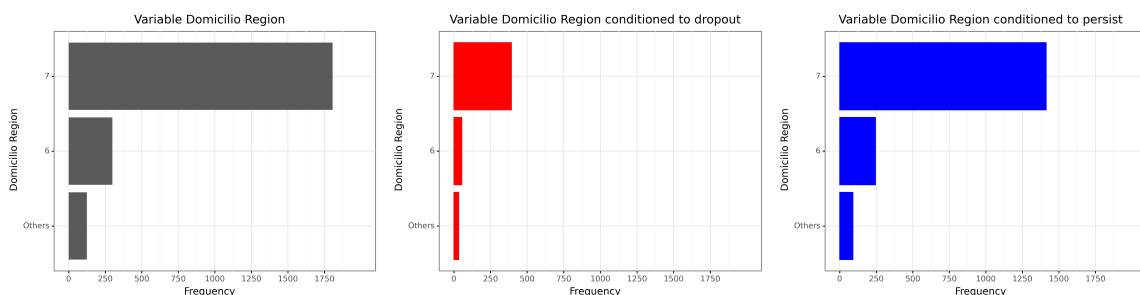


FIGURE B.18: (Conditional) Analysis distribution for Domicilio Region: Slightly higher dropout in Others, lower dropout in 6.

10. Carrera: Categorical variable showing the student program. It has 7 different values with engineering only. We can see a higher dropout in CIVIL EN COMPUTACIÓN, CIVIL DE MINAS, and CIVIL EN BIOINFORMÁTICA.

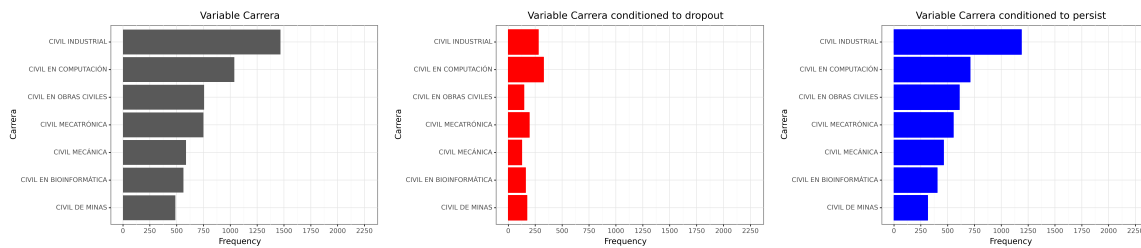


FIGURE B.19: (Conditional) Analysis distribution for Carrera: Higher dropout in CIVIL EN COMPUTACIÓN, CIVIL DE MINAS, and CIVIL EN BIOINFORMÁTICA.

11. Campus: Categorical variable describing the student's campus. It has 2 possible values: CURICO and TALCA. We can see that students studying in TALCA have higher probabilities of dropout, which could be significant.

TABLE B.7: Analysis distribution for Campus

| Campus | Total frequency | Total percent | Dropout frequency | Dropout percent | Persist frequency | Persist percent |
|--------|-----------------|---------------|-------------------|-----------------|-------------------|-----------------|
| CURICO | 2091            | 93.81%        | 430               | 89.58%          | 1661              | 94.97%          |
| TALCA  | 138             | 6.19%         | 50                | 10.42%          | 88                | 5.03%           |

12. Tipo Educacion: Categorical variable indicating the school education type of the student. It has 9 possible values. We can see that students with CIENTIFICO HUMANISTA - DIURNA education have lower probabilities of dropout.

TABLE B.8: Analysis distribution for Tipo Educacion: lower probability of dropout with CIENTIFICO HUMANISTA - DIURNA education.

| Education               | Total frequency | Total percent | Dropout frequency | Dropout percent | Persist frequency | Persist percent |
|-------------------------|-----------------|---------------|-------------------|-----------------|-------------------|-----------------|
| CIENTIF. HUMAN.-DIURNA  | 1718            | 77.07%        | 335               | 69.79%          | 1383              | 79.07%          |
| INDUSTRIAL              | 269             | 12.07%        | 67                | 13.96%          | 202               | 11.55%          |
| COMERCIAL I.            | 114             | 5.11%         | 36                | 7.5%            | 78                | 4.46%           |
| CIENTIF. HUMAN.-VESPER. | 29              | 1.3%          | 10                | 2.08%           | 19                | 1.09%           |
| SERVICIOS Y TECNICA     | 28              | 1.26%         | 10                | 2.08%           | 18                | 1.03%           |
| AGRICOLA                | 9               | 0.4%          | 2                 | 0.42%           | 7                 | 0.40%           |
| ALUMNOS LIBRES          | 4               | 0.18%         | 1                 | 0.21%           | 3                 | 0.17%           |
| ESTUDIOS EXTRANJEROS    | 1               | 0.04%         | 0                 | 0.0%            | 1                 | 0.06%           |
| Nulls                   | 57              | 2.56%         | 19                | 3.96%           | 38                | 2.17%           |

13. Dependencia: Categorical variable indicating the school type of the student. It has 3 possible values: PARTICULAR PAGADO (private school), PARTICULAR SUBVENCIONADO (private with state support), and MUNICIPAL (public school). We can see a similar behavior among persist, dropout, and total percent. Although it is not observed that affects dropout, it was considered that, since this variable was maintained in UAI's dataset, it would also be maintained here for future combined analyses.

TABLE B.9: Analysis distribution for Dependencia: similar behavior among persist, dropout and total percent.

| Education           | Total frequency | Total percent | Dropout frequency | Dropout percent | Persist frequency | Persist percent |
|---------------------|-----------------|---------------|-------------------|-----------------|-------------------|-----------------|
| PARTICULAR SUBVENC. | 1172            | 52.58%        | 251               | 52.29%          | 921               | 52.66%          |
| MUNICIPAL           | 875             | 39.26%        | 189               | 39.38%          | 686               | 39.22%          |
| PARTICULAR PAGADO   | 128             | 5.74%         | 24                | 5.0%            | 104               | 5.95%           |
| Nulls               | 54              | 2.42%         | 16                | 3.33%           | 38                | 2.17%           |

14. Grupo Ingreso: Categorical variable indicating the type of admission. It has 2 possible values: Ingreso PSU (regular admission) or Ingreso Especial (special admission, such as sport scholarship). We can see a similar behavior among persist, dropout, and total percent. Although it is not observed that affects dropout, it was considered that, since this variable was maintained in UAI's dataset, it would also be maintained here for future combined analyses.

TABLE B.10: Analysis distribution for Grupo Ingreso: Similar behavior among persist, dropout and total percent.

| Type of admission | Total frequency | Total percent | Dropout frequency | Dropout percent | Persist frequency | Persist percent |
|-------------------|-----------------|---------------|-------------------|-----------------|-------------------|-----------------|
| Ingreso PSU       | 2155            | 96.68%        | 457               | 95.21%          | 1698              | 97.08%          |
| Ingreso Especial  | 74              | 3.32%         | 23                | 4.79%           | 51                | 2.92%           |

15. Preferencia: Numerical variable indicating the preference of the students for Universidad de Talca in her/his application. It ranges from 1 to 10, where 1 means that Universidad de Talca was the first option. Although it can not be observed at simple sight, when comparing the proportions of the values, we observe a lower proportion of dropout in Preferencia 1.

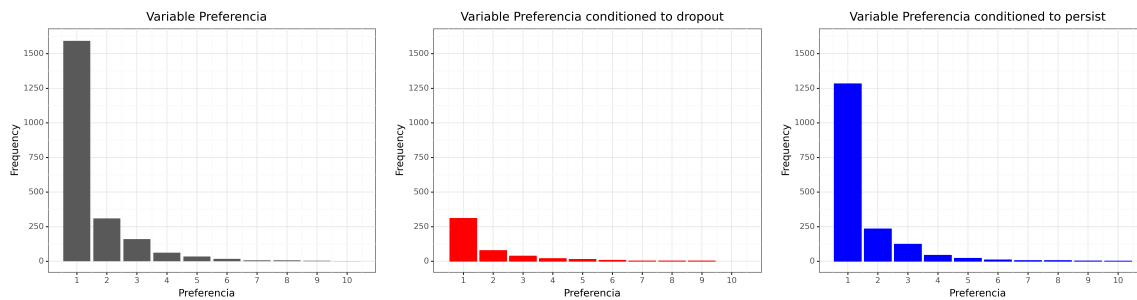


FIGURE B.20: (Conditional) Analysis distribution for Preferencia: Similar behavior among the proportion of the values in persist, dropout and total.

16. Matricula: Unique ID, in the datasets DBB and InscritasReprobadasPorAño. It was used to merge both datasets.
17. PTJE PPS: Numerical variable indicating the weighted student's university admission score. It ranges from 487.2 to 823.81. We can see that students who dropout have lower values than students who persist.

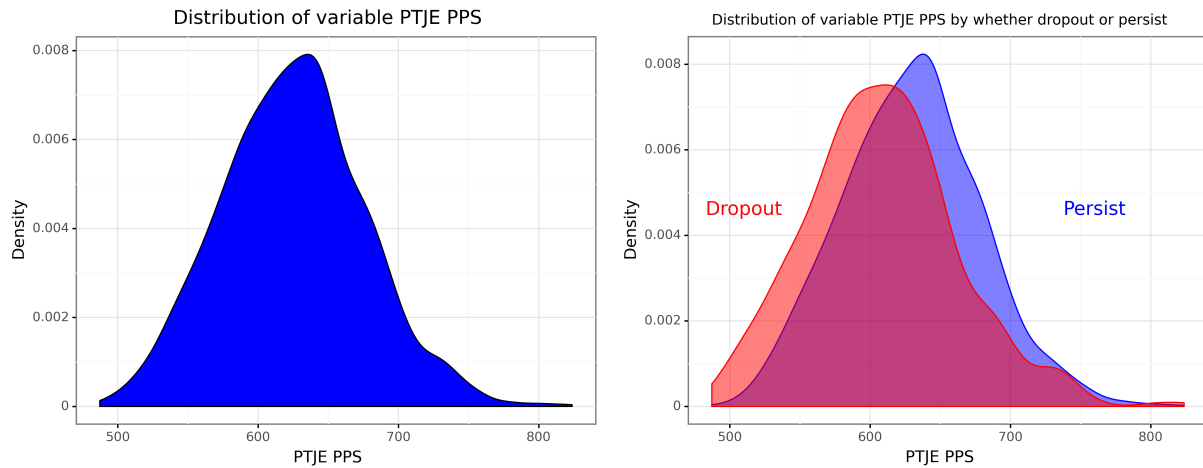


FIGURE B.21: (Conditional) Analysis distribution for PTJE PPS: lower values increases dropout.

18. PTJE LENG: Numerical variable indicating language test score. It ranges from 318 to 811. Although it was expected that students who dropout will have lower values than students who persist, this almost can not be appreciated.

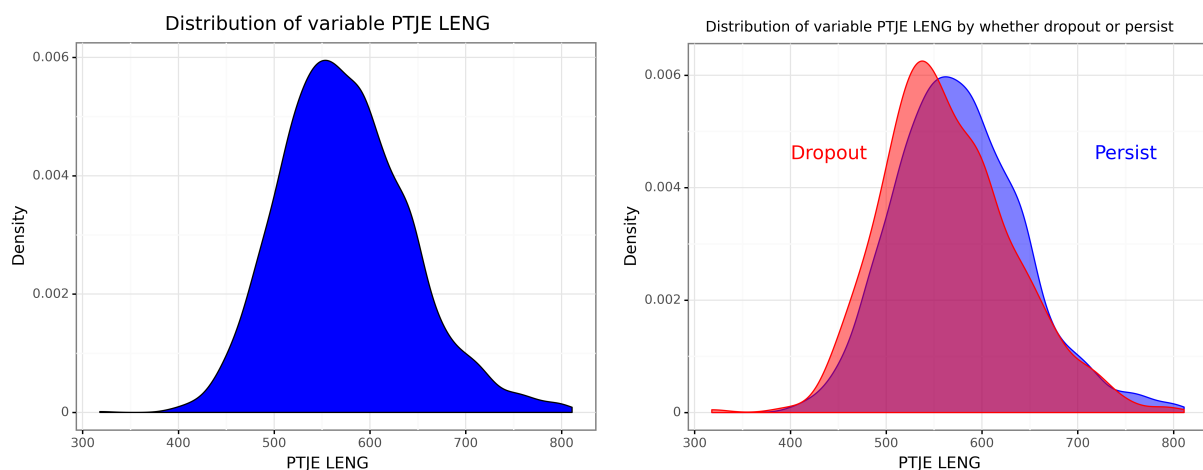


FIGURE B.22: (Conditional) Analysis distribution for PTJE LENG: similar behavior between distributions.

19. PTJE MAT: Numerical variable indicating math test score. It ranges from 312 to 850. We can see that students who dropout have lower values than students who persist.

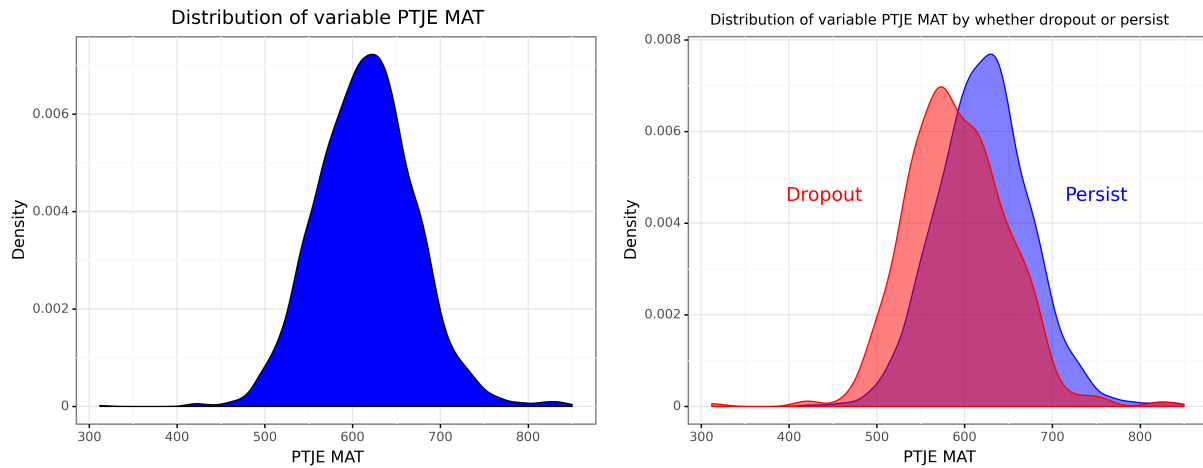


FIGURE B.23: (Conditional) Analysis distribution for PTJE MAT: lower values increases dropout.

20. PTJE CS: Numerical variable indicating science test score. It ranges from 300 to 817. We can see that students who dropout have lower values than students who persist.

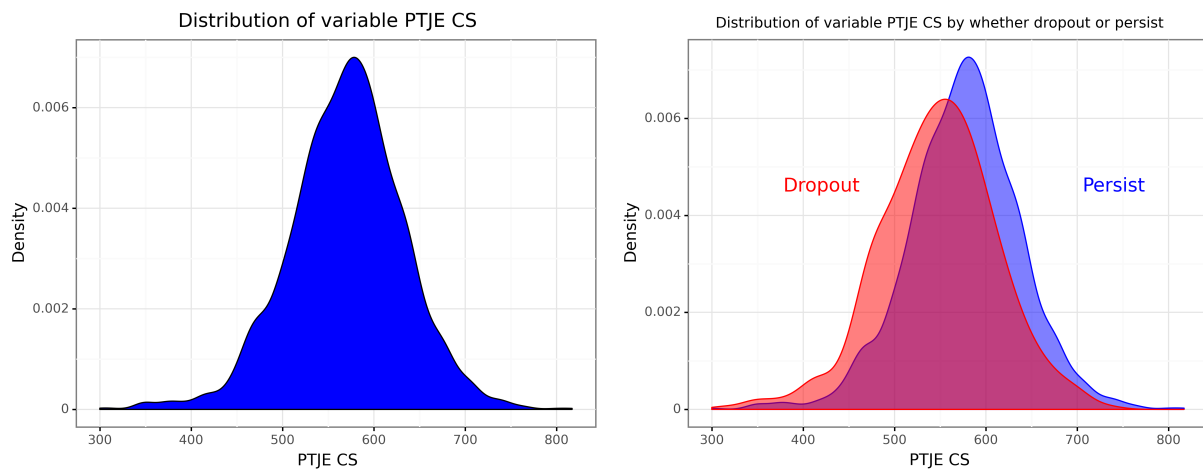


FIGURE B.24: (Conditional) Analysis distribution for PTJE CS: lower values increases dropout.

21. PTJE HISTORIA: Numerical variable indicating history test score. It ranges from 293 to 809. Although it was expected that students who dropout will have lower values than students who persist, this can not be appreciated.



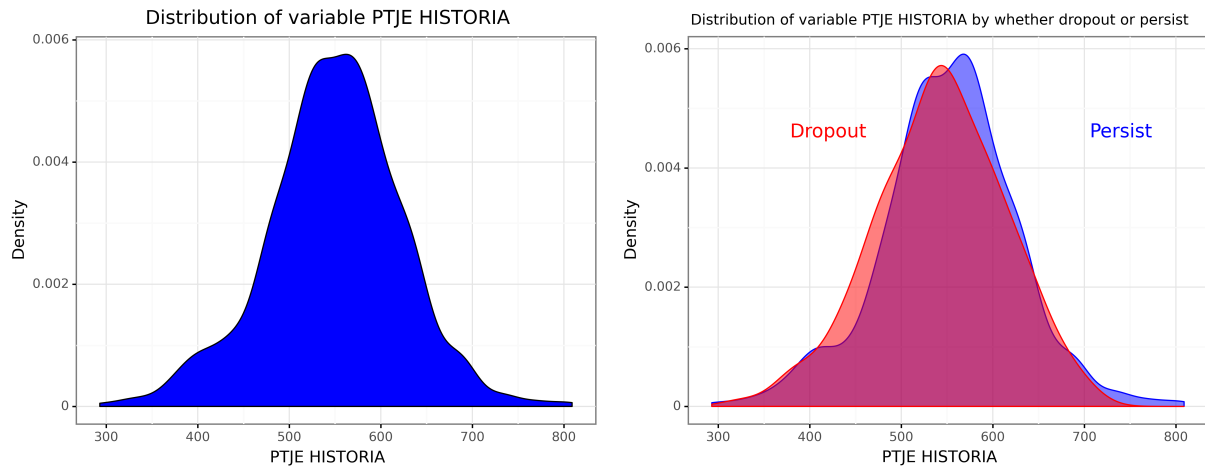


FIGURE B.25: (Conditional) Analysis distribution for PTJE HISTORIA: similar behavior between distributions.

22. PTJE NEM: Numerical variable that corresponds to the high school grades (average of the last 4 high school years, scaled to the range 0-850). It ranges from 404 to 825. We can see that students who dropout have lower values than students who persist.

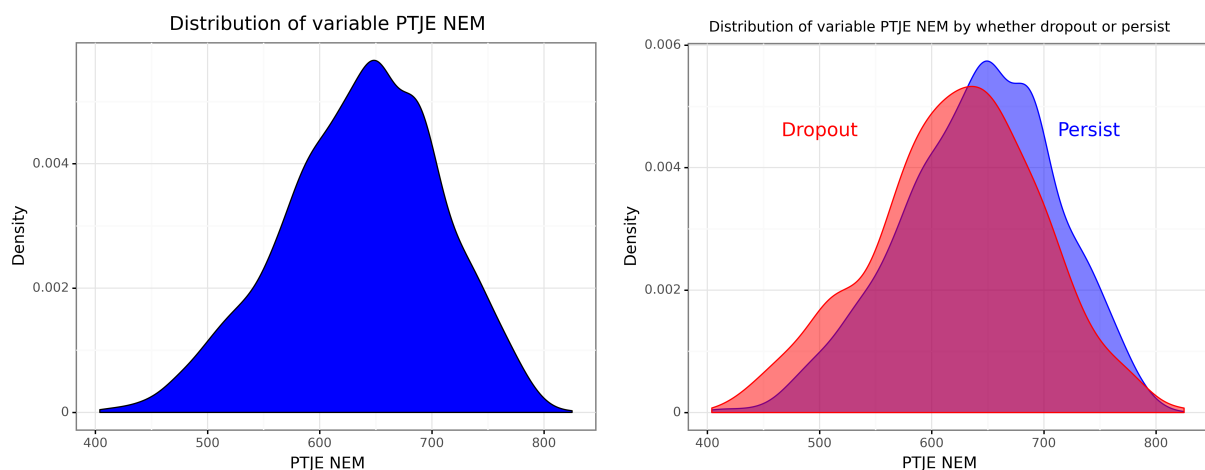


FIGURE B.26: (Conditional) Analysis distribution for PTJE NEM: lower values increases dropout.

23. PTJE RKG: Numerical variable indicating the student high school ranking, which is a selection factor that considers the academic performance of a student in relation to their educational context. It ranges from 404 to 850. We can see that students who dropout have lower values than students who persist.

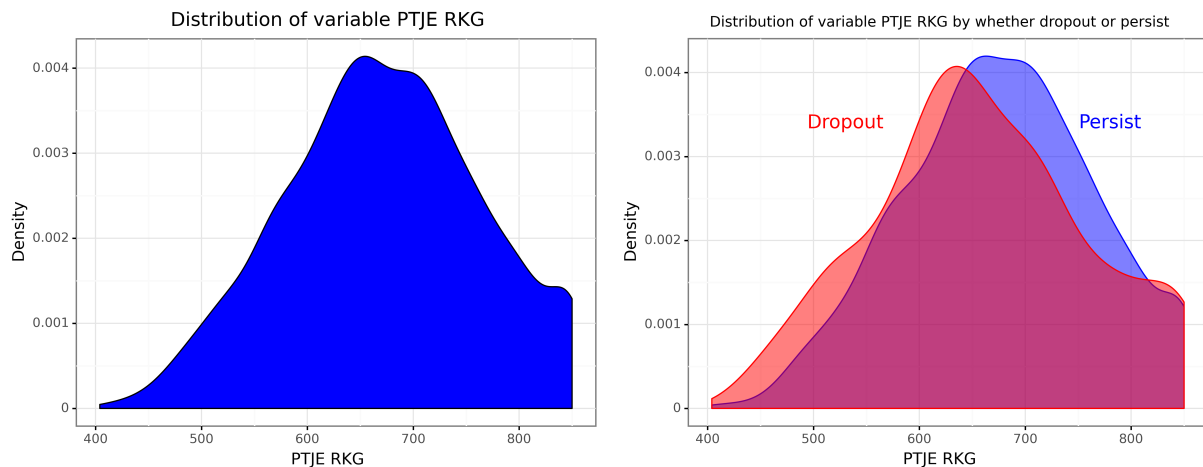


FIGURE B.27: (Conditional) Analysis distribution for PTJE RKG: lower values increases dropout.

24. Ingreso bruto familiar: Categorical ordinal variable that indicates the student's family income. It has 12 possible income ranges, distributed as shown in Table B.11. We can see a slightly higher probability of dropout in the lower incomes, and also in the higher income.

TABLE B.11: Analysis distribution for Ingreso bruto familiar: higher probability of dropout in lower incomes.

| Range                 | Total frequency | Total percent | Dropout frequency | Dropout percent | Persist frequency | Persist percent |
|-----------------------|-----------------|---------------|-------------------|-----------------|-------------------|-----------------|
| 0- 144.000            | 183             | 8.21%         | 54                | 11.25%          | 129               | 7.38%           |
| 144.001 - 288.000     | 695             | 31.18%        | 158               | 32.92%          | 537               | 30.70%          |
| 288.001 - 432.000     | 466             | 20.91%        | 96                | 20.0 %          | 370               | 21.15%          |
| 432.000 - 576.001     | 251             | 11.26%        | 57                | 11.88%          | 194               | 11.09%          |
| 576.001 - 720.000     | 167             | 7.49%         | 30                | 6.25%           | 137               | 7.83%           |
| 720.001 - 864.000     | 134             | 6.01%         | 23                | 4.79%           | 111               | 6.35%           |
| 864.001 - 1.008.000   | 99              | 4.44%         | 20                | 4.17%           | 79                | 4.52%           |
| 1.008.001 - 1.152.000 | 53              | 2.38%         | 6                 | 1.25%           | 47                | 2.69%           |
| 1.152.001 - 1.296.000 | 43              | 1.93%         | 8                 | 1.67%           | 35                | 2.00%           |
| 1.296.001 - 1.440.000 | 26              | 1.17%         | 1                 | 0.21%           | 25                | 1.43%           |
| 1.440.001 - 1.584.000 | 21              | 0.94%         | 1                 | 0.21%           | 20                | 1.14%           |
| 1.584.001 o más       | 74              | 3.32%         | 23                | 4.79%           | 51                | 2.92%           |
| Nulls                 | 17              | 0.76%         | 3                 | 0.62%           | 14                | 0.80%           |

25. Educ padres: Categorical variable that indicates if the student's parents are professionals. It has 3 possible values: Alguno Profesional (one of them is professional), Ambos Profesionales (both are professionals), and Ninguno Profesional (none of them is professional) distributed as shown in Table B.12. We can see a slightly higher probability of dropout when none of the parents are professionals.

TABLE B.12: Analysis distribution for Educ padres: higher probability of dropout when none of the parents are professionals.

| Education           | Total frequency | Total percent | Dropout frequency | Dropout percent | Persist frequency | Persist percent |
|---------------------|-----------------|---------------|-------------------|-----------------|-------------------|-----------------|
| Ninguno Profesional | 1747            | 78.38%        | 385               | 80.21%          | 1362              | 77.87%          |
| Alguno Profesional  | 271             | 12.16%        | 58                | 12.08%          | 213               | 12.18%          |
| Ambos Profesionales | 94              | 4.22%         | 17                | 3.54%           | 77                | 4.40%           |
| Nulls               | 117             | 5.25%         | 20                | 4.17%           | 97                | 5.55%           |

26. **Semestres:** Categorical ordinal variable indicating the semester when the student drop out of university. Being 1 in case the student has dropped out in the first year and 0 in other cases.

TABLE B.13: Analysis distribution for Semestres

| Semester | Frequency |
|----------|-----------|
| 0        | 1729      |
| 1        | 472       |

## Appendix C

# Selection of parameters and variables

This section describes the method used for the selection of hyperparameters and variables for each model.

### C.1 K-Nearest Neighbors (KNN)

For KNN we considered the numerical variables (mat, lang, pps, nem, optional, and ranking) and the parameter  $K$ . Ordinal and nominal variables were discarded by the use of the euclidean distance. For the both-universities-dataset, the selected variables were mat, lang, pps, and optional. Figure C.1 shows the results for the  $F_1$  score for the  $K$  values from 1 to 60. Over  $K = 25$  the  $F_1$  score begins to stabilize, so any high  $F_1$  score  $K$  value could be used. We used  $K = 29$ , since it has a high  $F_1$  score, and keeping a lower  $K$  value gives more flexibility to the model. A higher  $K$  value would produce a less flexible model, leading to underfitting.

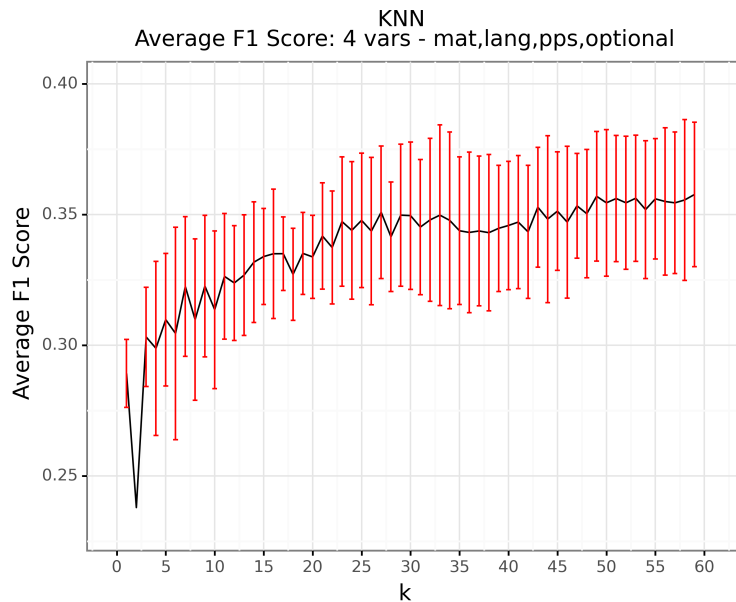


FIGURE C.1: KNN results. Both universities.

For Universidad Adolfo Ibáñez, the selected variables were mat, nem, optional, and pps. Figure C.2 shows the results for the  $F_1$  score for the  $K$  values from 1 to 60. Over  $K = 17$  the  $F_1$  score begins to stabilize until a small drop over  $K = 26$ . We used  $K = 17$ , since keeping a lower  $K$  value gives more flexibility to the model.

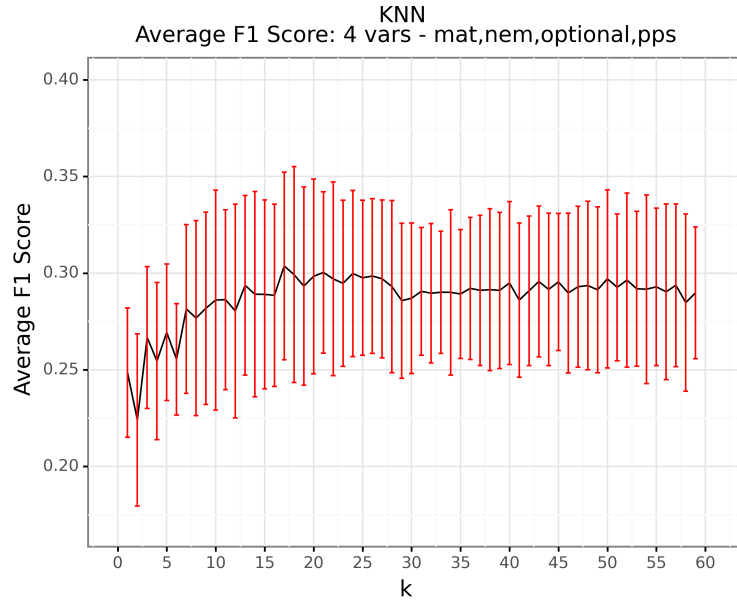


FIGURE C.2: KNN results. UAI.

For Universidad de Talca, the selected variables were mat, ranking, and optional. It was noted a rising slope with increasing  $K$  value, this stabilized on  $K = 70$ , before a small decrease. We used  $K = 71$  as the best result, keeping the value far from the drop.

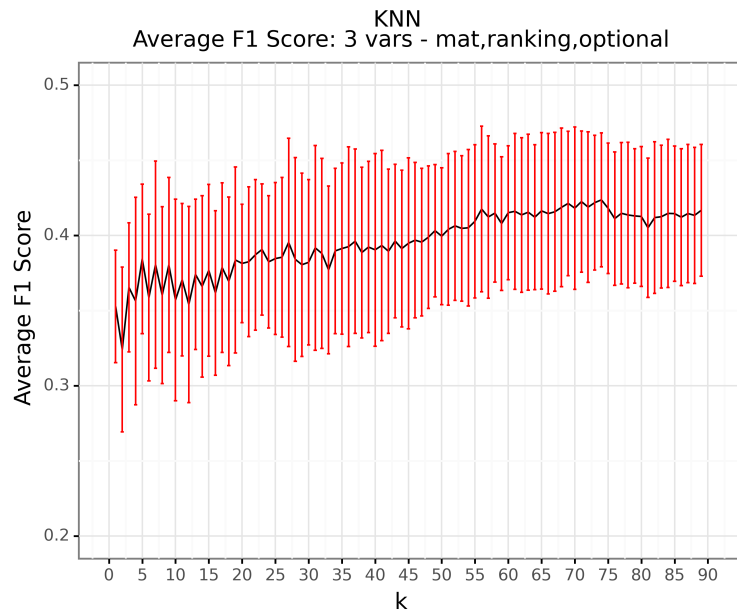


FIGURE C.3: KNN results. U de Talca.

## C.2 Support vector machine (SVM)

For the support vector machine model, we considered the numerical variables (mat, lang, pps, nem, optional, and ranking), the penalization  $C$  parameter, and kernels. Ordinal and nominal variables were discarded by the use of the euclidean distance. We considered  $C \in \{0.1, 1, 10, 100\}$  and the

following kernels: polynomial (poly), radial basis function (rbf), and sigmoid. All possible combinations of the kernels, variables, and  $C$  values were done. For the both-universities-dataset, we obtained the best scores using all variables. Figure C.4 shows the results for the kernels applied, where poly gives the best scores. Although, we can see that all results for the poly kernel are similar,  $C = 10$  has a smaller deviation, so it was considered the best result.

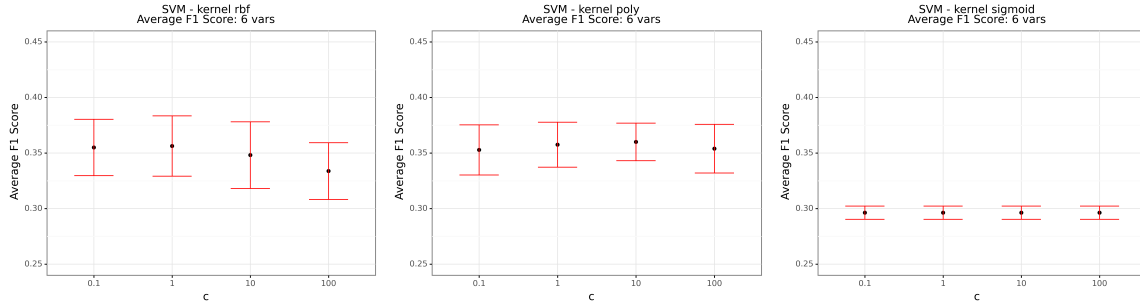


FIGURE C.4: SVM results. Both universities.

For Universidad Adolfo Ibáñez, the best scores were obtained using mat, nem, lang, and optional, and the polynomial kernel. Figure C.5 shows that, even though all results for the poly kernel are statistically similar,  $C = 1$  has a higher average score, so it was considered the best result.

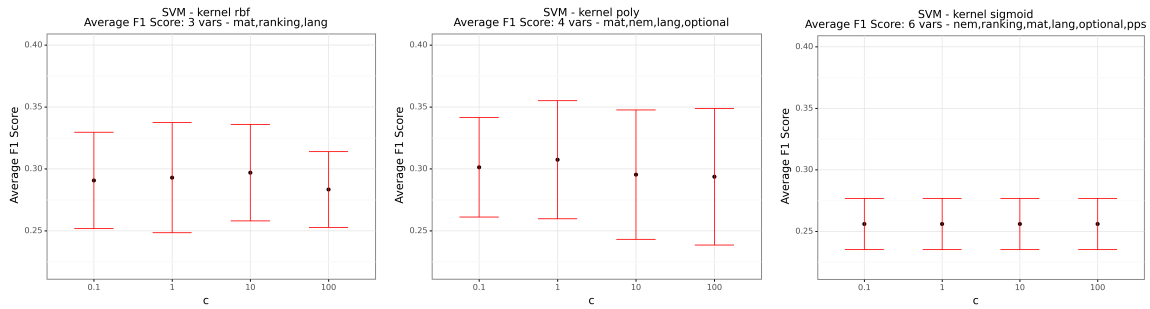


FIGURE C.5: SVM results. UAI.

For Universidad de Talca, best scores were obtained using mat, nem, ranking, optional, and lang, and the rbf kernel. However, the second best score, using lang, mat, and nem, and the polynomial kernel, showed a slightly lower deviation, therefore, it was considered the best result with  $C = 1$ .

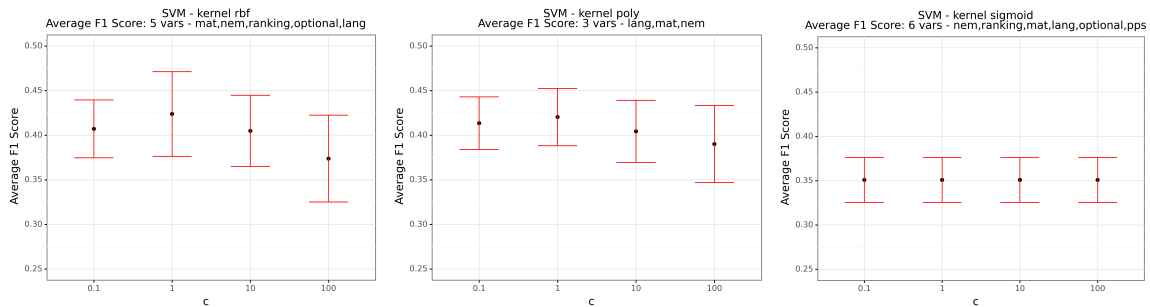


FIGURE C.6: SVM results. U de Talca.

### C.3 Decision tree

For the decision tree algorithm, we used almost all the variables (gender, school, admission, nem, ranking, mat, lang, optional, pps, preference, commune, region, and university) with exception of ID and year. This selection was applied as a decision tree selects the best variables in their training process. However, the implementation of the algorithm can not handle categorical variables properly, therefore one hot encoding was used on them. For the parameters, we considered the maximum depth of the tree (max depth), the minimum number of samples required to split an internal node (min samples split), and the minimum number of samples required to be at a leaf node (min samples leaf). For the both-universities-dataset, it was noticed that the min samples leaf value strongly affected the performance of the model, without obtaining significant variations of the score in combinations with the other parameters (using no max depth and min samples split equal 2). Figure C.7 shows min samples leaf values from 150 to 200, where minimal differences can be appreciated, which could be significant. In spite of that, 187 was selected as the best result, since it has a higher average value, even though it could not be significant.

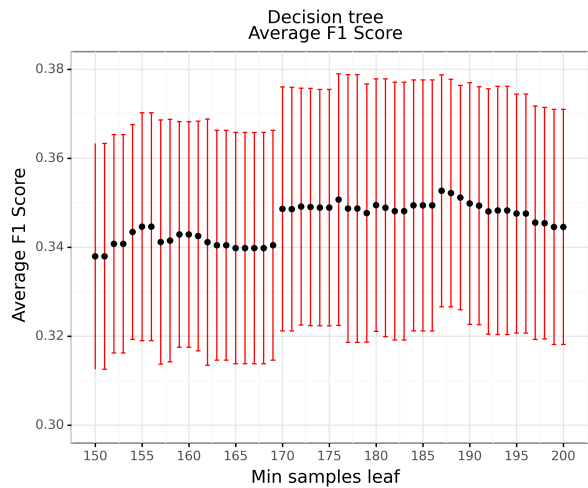


FIGURE C.7: Decision tree results. Both universities.

For Universidad Adolfo Ibáñez, the variable university was not used since it has only one value. As in the both-universities-dataset, it was noticed that the min samples leaf value strongly affected the performance of the model, without obtaining significant variations of the score in combinations with the other parameters (using no max depth and min samples split equal 2). Figure C.8 shows min samples leaf values from 1 to 90, showing slightly higher values in values 28, 47, 48, and 49. Despite the minimal differences, 48 was selected as the best result, since it has a lower deviation.

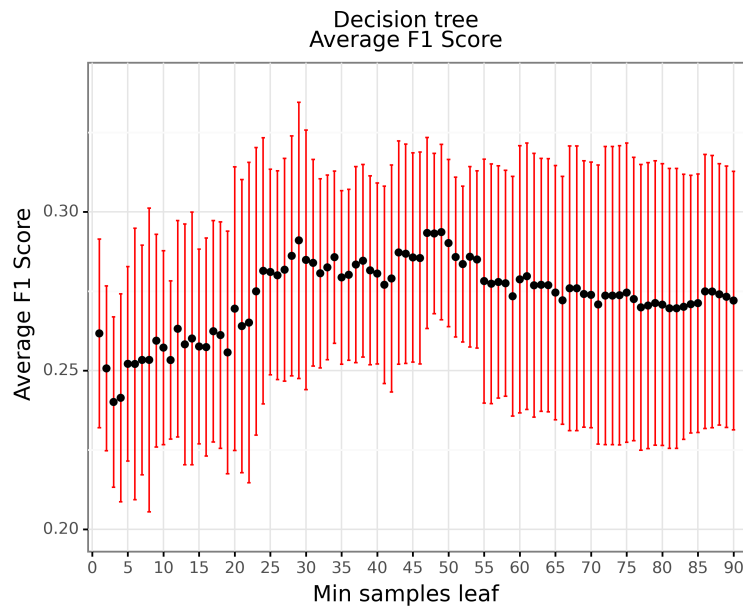


FIGURE C.8: Decision tree results. UAI.

For Universidad de Talca, once again, the min samples leaf value strongly affected the performance of the model, without obtaining significant variations of the score in combinations with the other parameters (using no max depth and min samples split equal 2). Figure C.9 shows min samples leaf values from 90 to 200, showing the highest scores between 120 and 125. Min sample leaf values above 200 were also calculated, but showed no higher scores; then 123 Min sample leaf was selected as the best result, even though any value between 120 and 125 could be selected.

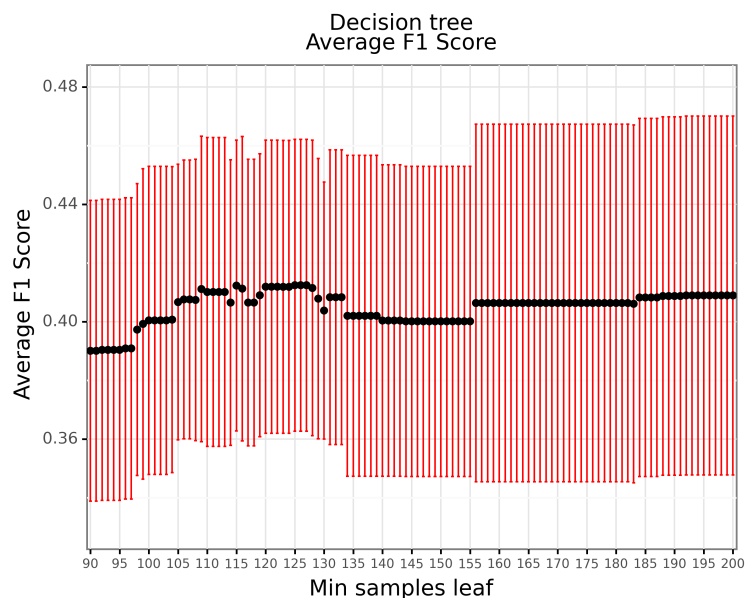


FIGURE C.9: Decision tree results. U de Talca.



## C.4 Random forest

Recall that Random Forest is an ensemble of decision trees. For this reason, we used the same variables and one hot encoding than decision trees. However, one hot encoding could be a problem in the application of the model, since one hot encoded variables that refer to a single categorical variable is separated in different trees. For the parameters, we considered the minimum number of samples required to be at a leaf node (min samples leaf), the number of trees used (n trees), and the max features used in each tree (max features). For the both-universities-dataset, unlike in the decision tree, all selected parameters affected the performance, so every possible combination using min samples leaf  $\in \{100, 150, 168, 180\}$ , number of trees  $\in \{10, 100, 200, 500\}$ , and max features  $\in \{3, 6, 10, 20\}$  were calculated. The highest average was achieved using 500 trees, 20 features, and min samples leaf of 100, as shown in figure C.10, however, it is not significantly better than other similar parameters.

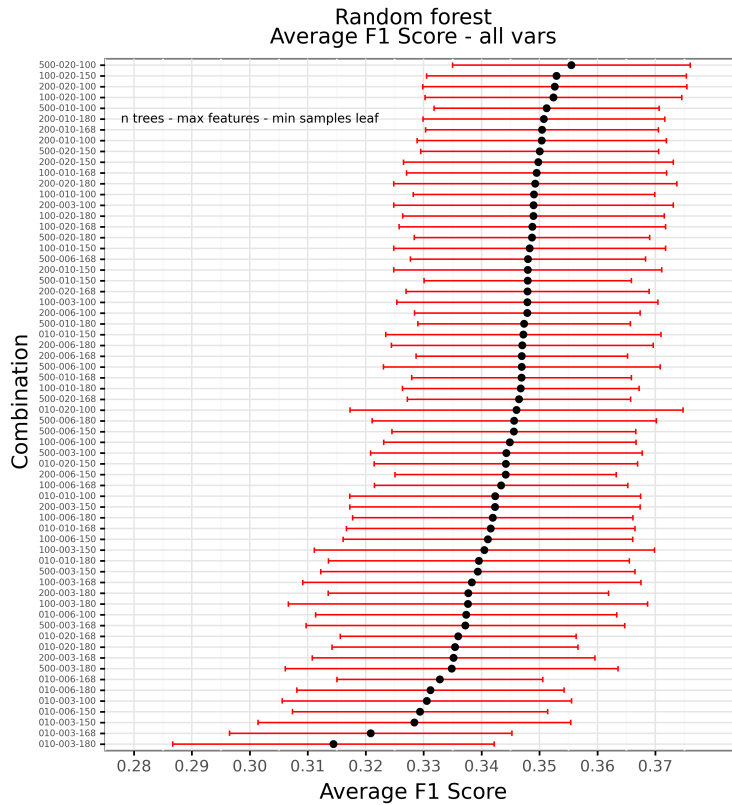


FIGURE C.10: Random forest results. Both universities.

For Universidad Adolfo Ibáñez, the variable university was not used since it has only one value. Every possible combination using min samples leaf  $\in \{20, 40, 60, 80\}$ , number of trees  $\in \{50, 200, 500\}$ , and max features  $\in \{6, 10, 15, 20\}$  were calculated. The highest average was achieved using 50 trees, 15 features, and min samples leaf of 20, as shown in figure C.11, however, it is not significantly better than other similar parameters.

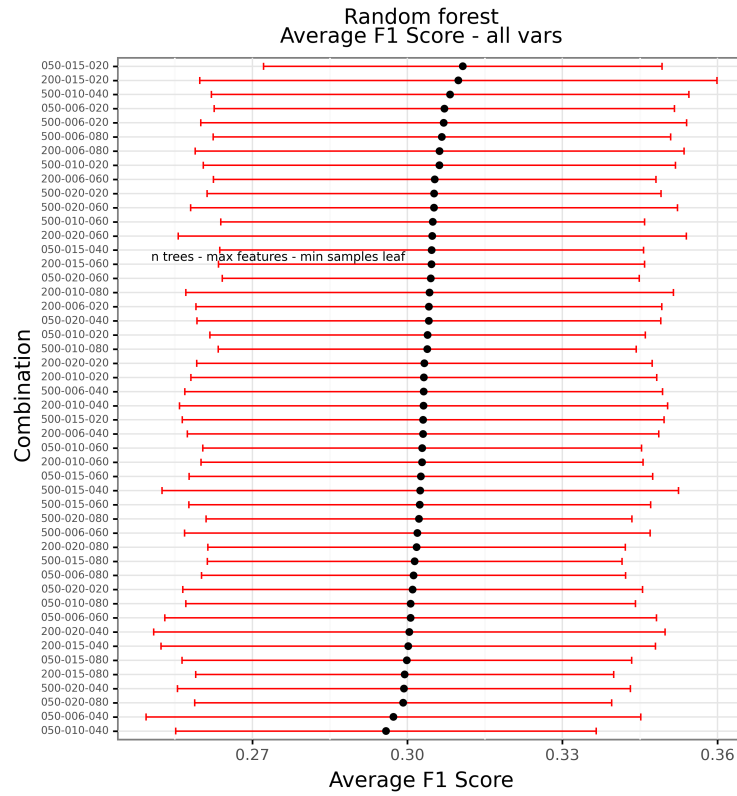


FIGURE C.11: Random forest results. UAI.

For Universidad de Talca, every possible combination using min samples leaf  $\in \{80, 120, 150\}$ , number of trees  $\in \{50, 200, 500\}$ , and max features  $\in \{6, 10, 15, 20\}$  were calculated. The highest average was achieved using, 500 trees, 20 features, and min samples leaf of 80, as shown in figure C.12, however, the second best average, used only 50 trees and had a slightly lower deviation, so the combination of 50 trees, 15 features, and min samples leaf of 150 was selected as the best result.

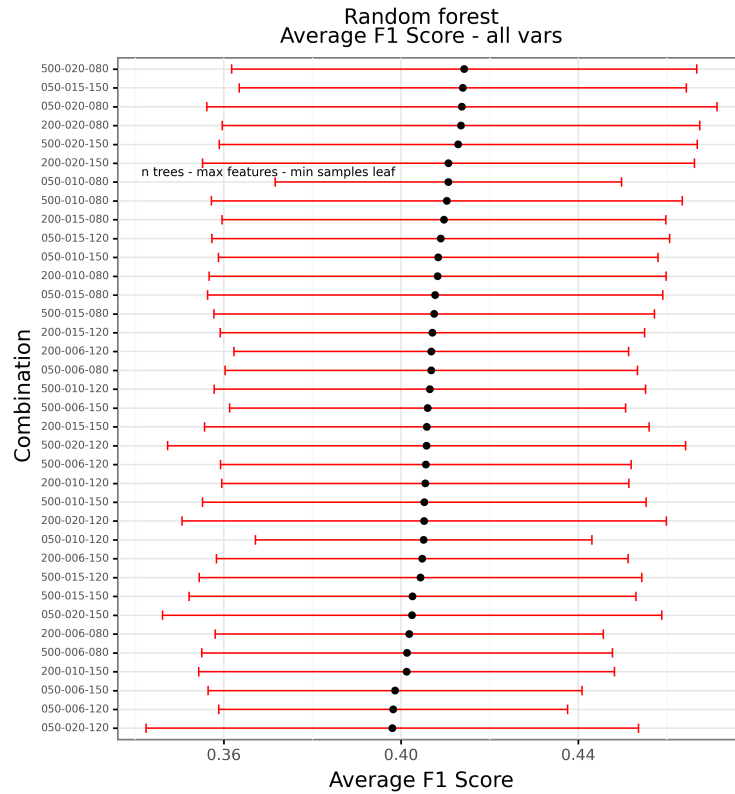


FIGURE C.12: Random forest results. U de Talca.

## C.5 Gradient boosting decision tree

Gradient tree boosting is, like random forest, an ensemble of decision trees. For this reason, we used the same variables and one hot encoding than decision tree and random forest. For the parameters, we considered the max depth, min samples split, min samples leaf (from decision tree), the number of trees used (n trees), the max features used in each tree (max features), and the learning rate. For the both-universities-dataset, it was first noticed that a low quantity of trees (lower than 200) improved the performance, that min samples split had no important effect, and that leaving max depth and learning rate in the default value (3 and 0.1) showed the best performance. Possible combinations similar to the previous models were tested. The second higher result was achieved using 110 trees, max features of 6, min samples leaf of 150, and leaving max depth, learning rate, and min samples split in their default values of 3, 0.1, and 2, as shown in figure C.13. This was considered the best result since it has a slightly lower deviation.

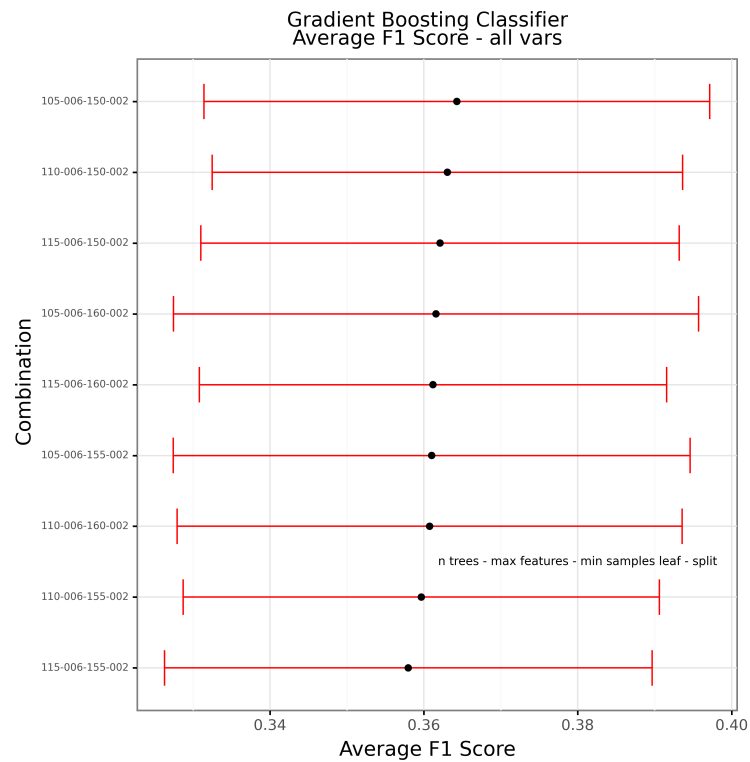


FIGURE C.13: Gradient tree boosting results. Both universities.

For Universidad Adolfo Ibáñez, again, the variable university was not used since it has only one value. It was noticed that a low quantity of trees (lower than 100) improved the performance. Again, max depth and learning rate showed best results when using the default value (3 and 0.1). The higher results are shown in figure C.14, where we selected the combination of 80 trees, max features of 20, min samples leaf of 60, and leaving max depth, learning rate, and min samples split in their default values of 3, 0.1, and 2. This combination had the second higher average score, but showed a slightly lower deviation.

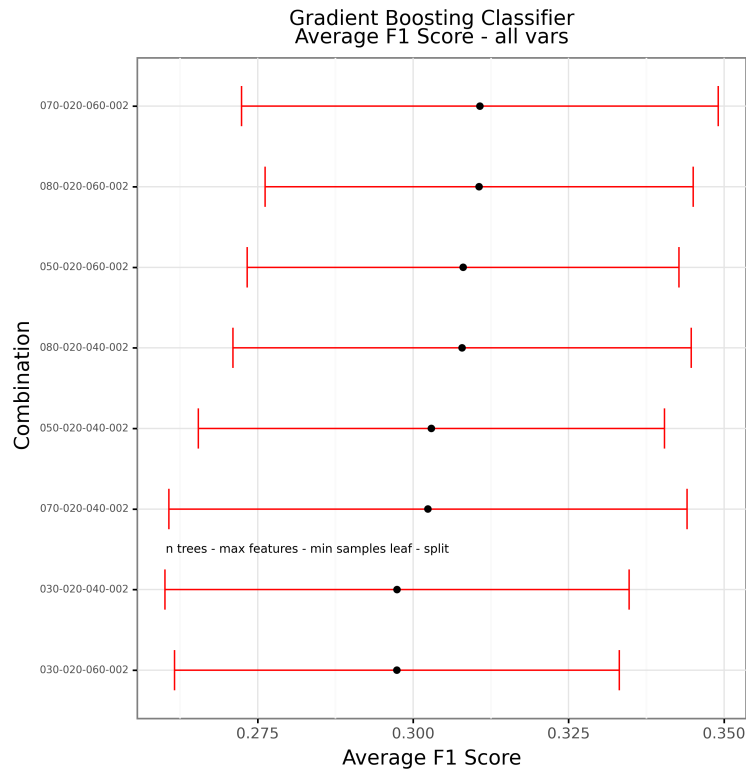


FIGURE C.14: Gradient tree boosting results. UAI.

For Universidad de Talca, again, the variable university was not used since it has only one value. In this dataset, it was noticed that a max depth of 10 improved the score, while a higher depth kept the score constant. The learning rate showed best results when using the default value (0.1), in the same way as the min samples split, which showed no visible effect when modified. The higher results are shown in figure C.15, where we selected the combination of 50 trees, max features of 15, min samples leaf of 130, max depth of 10, and leaving learning rate and min samples split in their default values of 0.1, and 2. This combination had the second higher average score, but showed a slightly lower deviation.

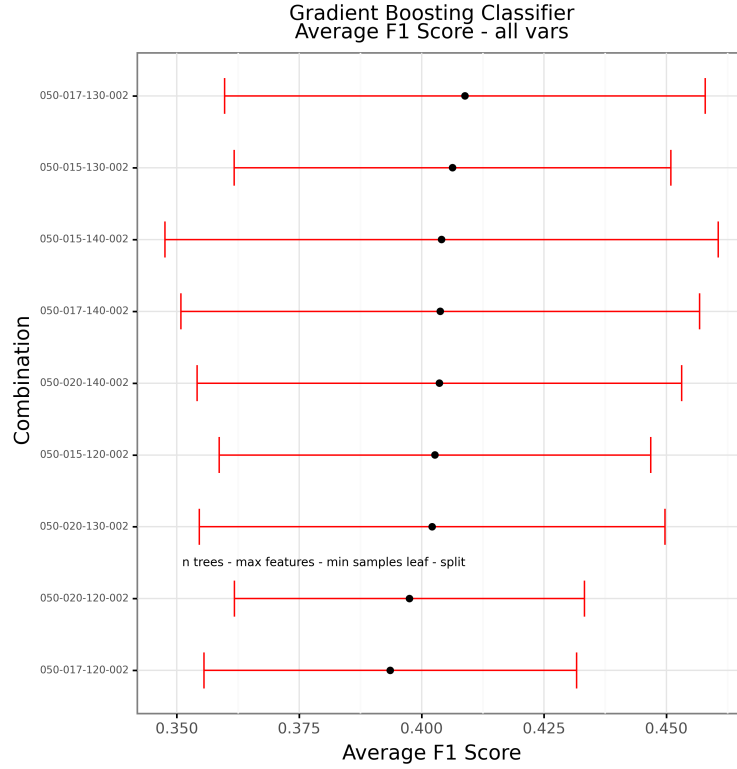


FIGURE C.15: Gradient tree boosting results. U de Talca.

## C.6 Naive Bayes (NB)

For the naive Bayes algorithm, the only parameter to modify was the Laplace smoothing. However, different implementations of the algorithm were used: Gaussian NB, multinomial NB, Complement NB, and Categorical NB, which differ in the assumptions made regarding the distribution of the values. While Gaussian NB is for numerical variables (mat, lang, pps, nem, optional, and ranking), others implementations are for categorical variables. These implementations differ in the scale of the Laplace smoothing parameter, therefore, we had to use different values. Specifically, for the Gaussian NB, we used values  $\in \{0, 1 \times 10^{-9}, 1 \times 10^{-5}, 1 \times 10^{-1}\}$ , while for other implementations we used  $\in \{0, 1, 10, 100\}$ . All combinations for these values and implementations were done, obtaining, for the both-universities-dataset, the best result for GaussianNB with the variables mat, pps, optional, ranking, and nem, independently of the smoothing parameter, as shown in figure C.16.

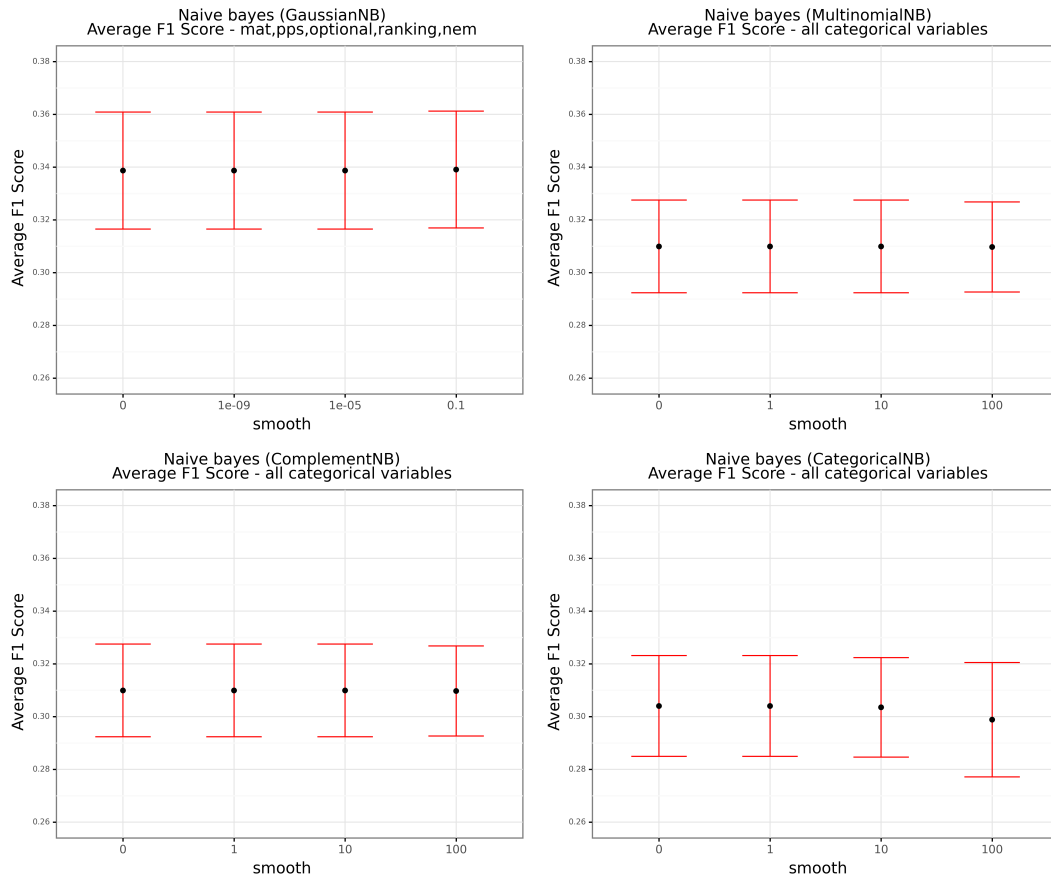


FIGURE C.16: Naive bayes results. Both universities.

For Universidad Adolfo Ibáñez, the best results were obtained for GaussianNB with the variables mat, optional, lang, and ranking, with slightly better results for a smoothing of 0.1, as shown in figure C.17, even though it is not significantly better than the other smoothing values.

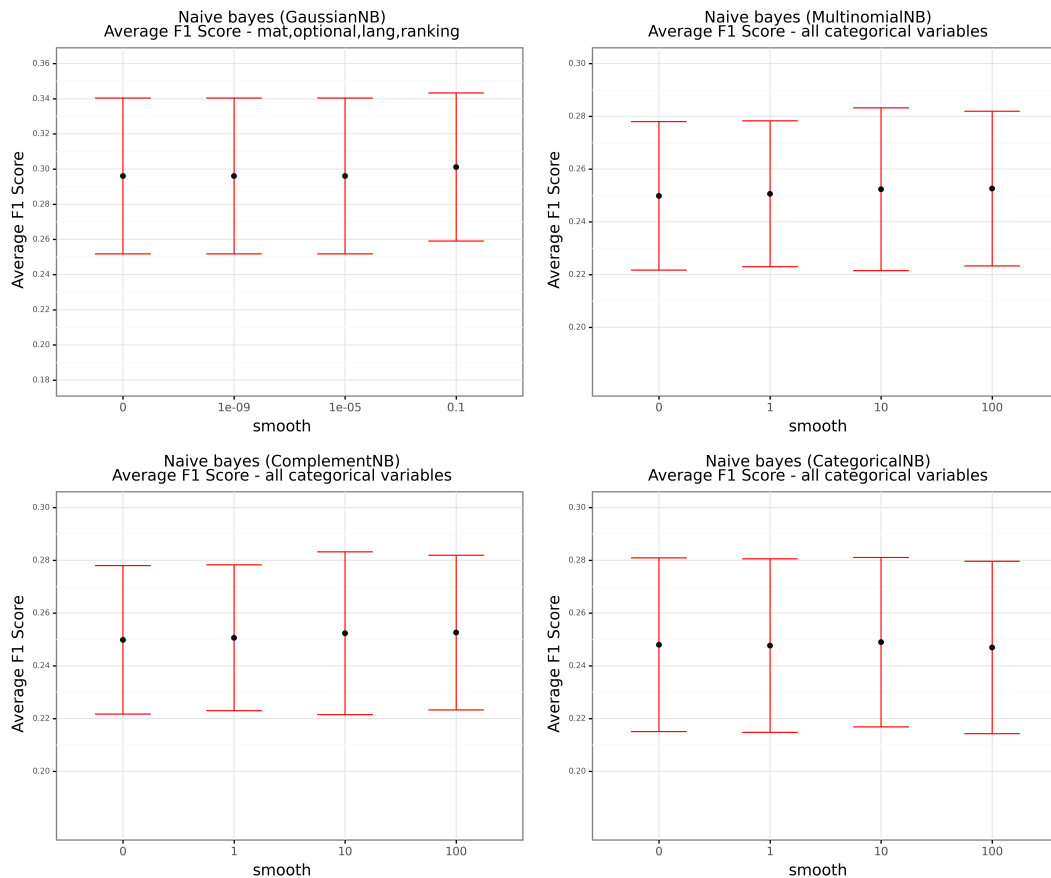


FIGURE C.17: Naive bayes results. UAI.

For Universidad de Talca, once again, the best results were obtained for GaussianNB, surprisingly with only two variables: mat and nem, with slightly better results for a smoothing lower than 0.1, as shown in figure C.18, even though it is not significantly better than the other smoothing values.



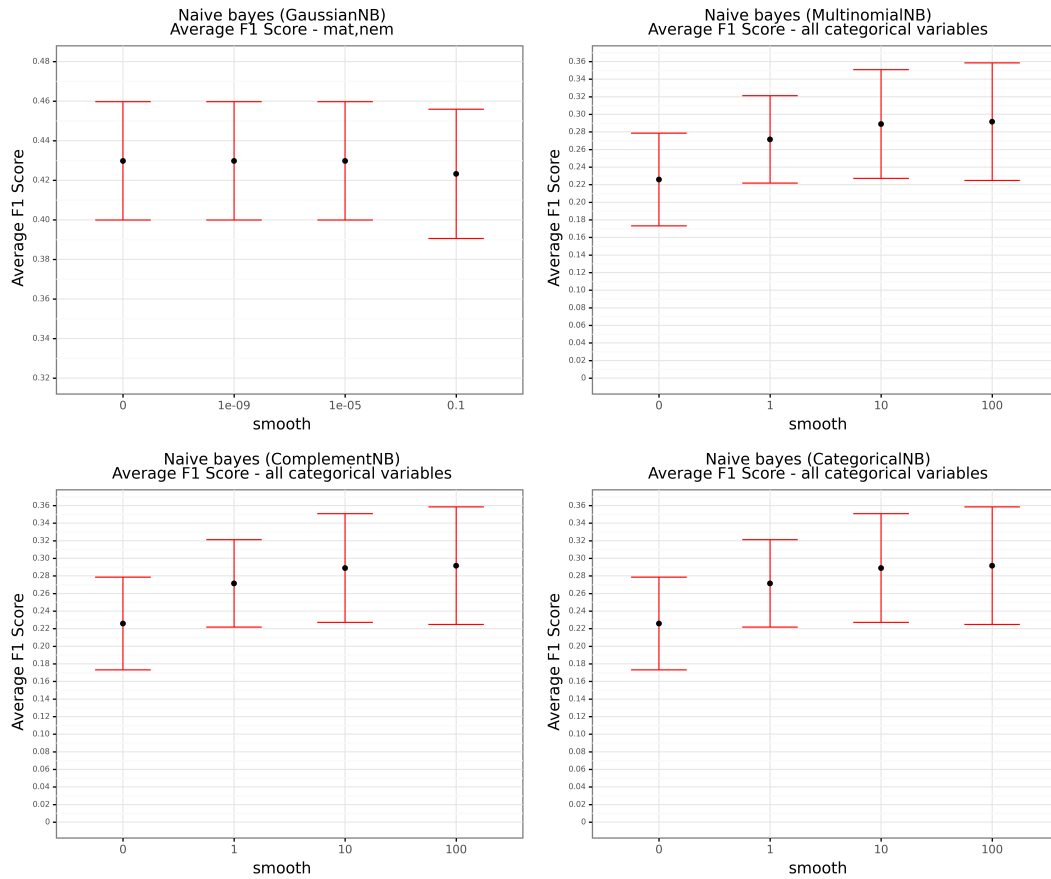


FIGURE C.18: Naive bayes results. U de Talca.

## C.7 Logistic regression

For the logistic regression model, it was necessary to define an algorithm for the optimization, and a regularization penalty to use. We considered the algorithm “Broyden–Fletcher–Goldfarb–Shanno” in the optimization problem, and a “l2” regularization penalty. All the variables were considered for the forward selection, with exception of ID and year, using one hot encoding in the categorical variables. For the both-universities-dataset, given the performance, we selected 7 variables: lang, pps, mat, optional, gender, school, admission, as shown in figure C.19, even though it is not significantly better than the other combinations.

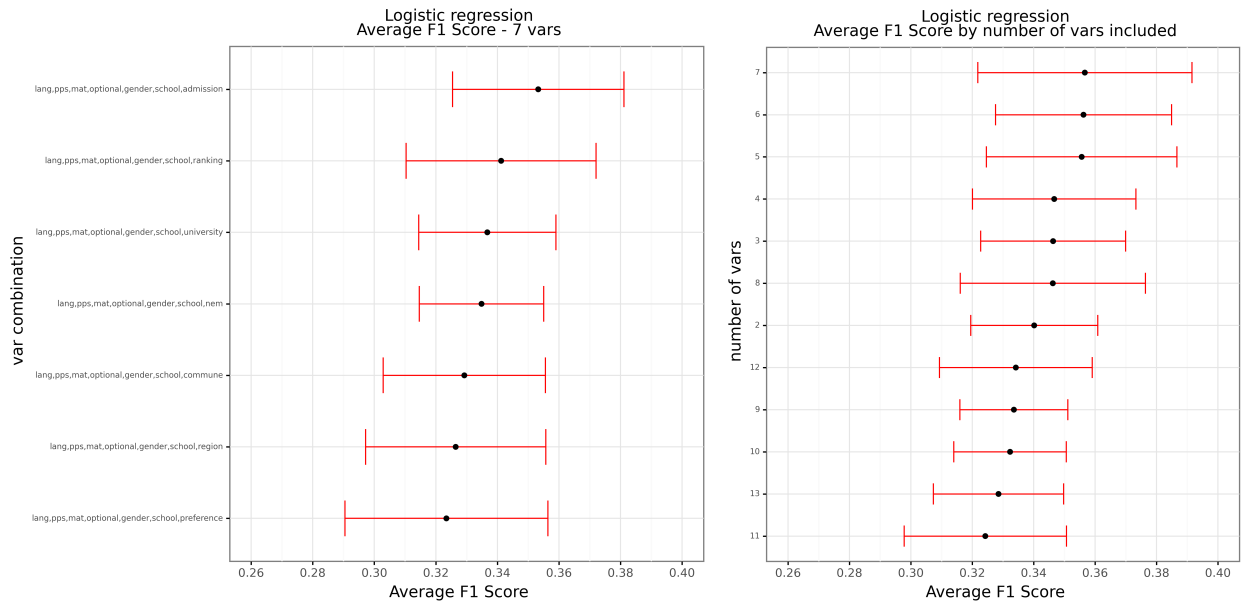


FIGURE C.19: Logistic regression results. Both universities.

For Universidad Adolfo Ibáñez, the best result was obtained using 4 variables: pps, lang, region, and gender, as shown in figure C.20, even though it is not significantly better than the other combinations.

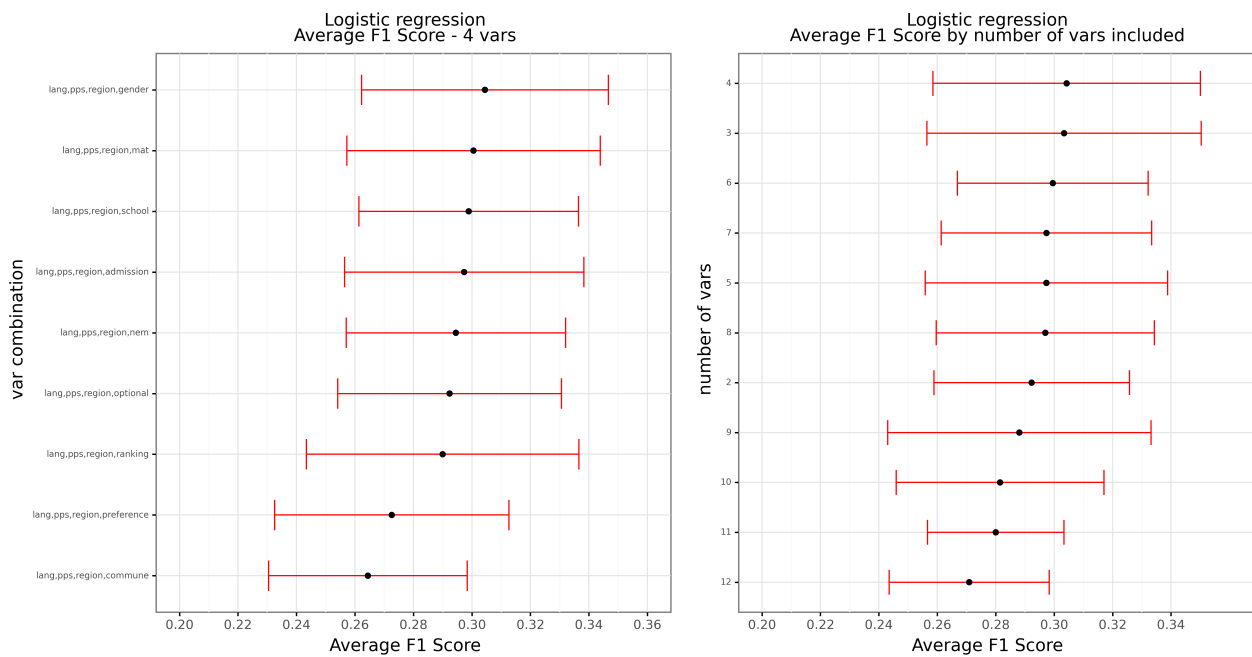


FIGURE C.20: Logistic regression results. UAI.

For Universidad de Talca, the higher score was obtained using 5 variables, however, the result using 6 variables obtained a similar average score, with a lower deviation. Therefore, the result of using the variables: pps, mat, nem, preference, admission, and lang was considered the best result.

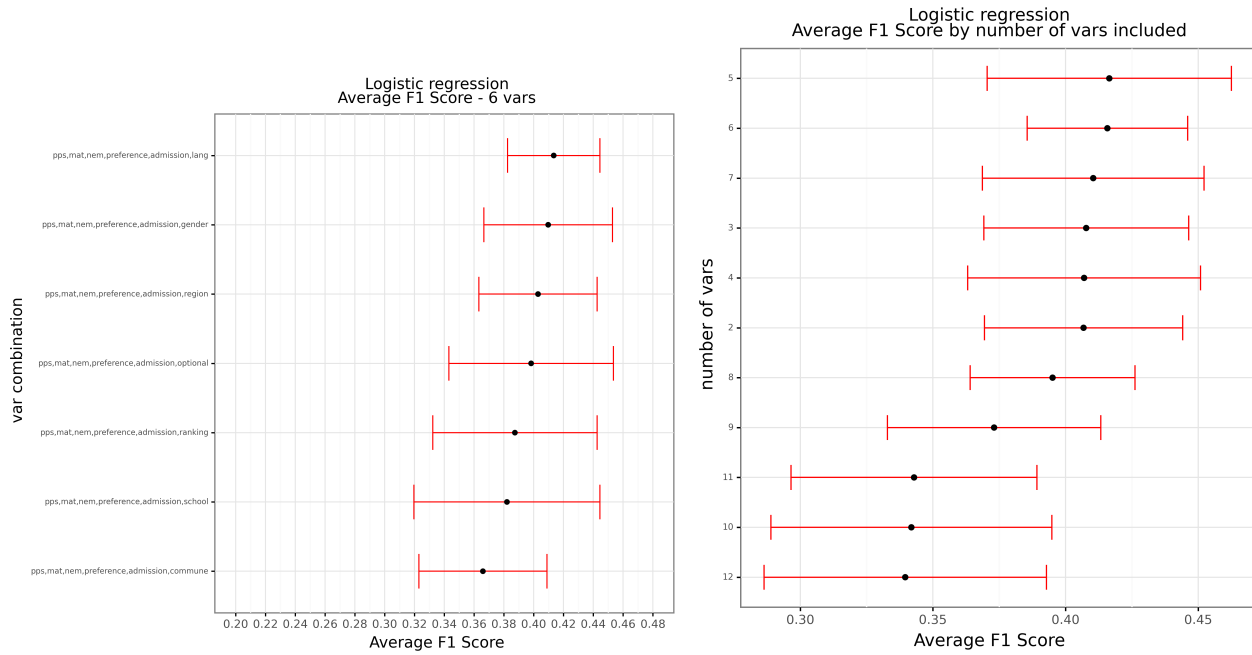


FIGURE C.21: Logistic regression results. U de Talca.

## C.8 Neural network

For the neural network model, we used a feedforward neural network with activation function ReLU for the hidden layers and Sigmoid for the output layer. It was intended to use a loss based on  $F_1$  score, but the implementation did not allow it. Because of this, we used binary crossentropy as loss function, and the algorithm “adam” in the optimization problem. We tuned the number of hidden layers and the number of neurons per layer. We considered initially all the variables, using one hot encoding in the categorical variables. Once multiple combinations were tested, for the both-universities-dataset, it was observed that a low number of layers (less than 20) produces results with less variance a slightly better average. After discarding architectures with a large number of layers, we proceed to search for the best set of variables. The variables selected were: lang, pps, gender, mat, school, optional, and admission. The combination of 2 layers, with 15 neurons per layer was selected as the best result, even though it is not significantly better than other combinations.

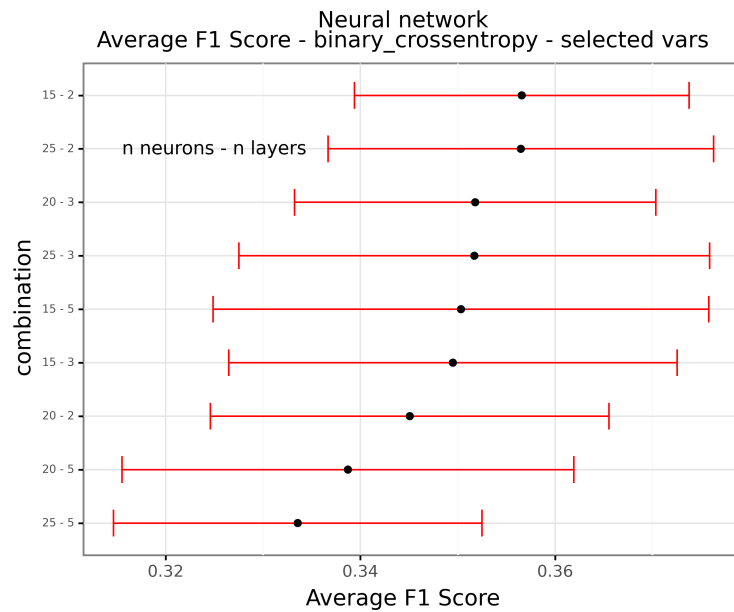


FIGURE C.22: Neural network results. Both universities.

For Universidad Adolfo Ibáñez, it was also observed that a low number of layers (less than 20) produces results with less variance. After discarding architectures with a large number of layers, we proceed to search for the best set of variables. The variables selected were: lang, pps, gender, region, and ranking. The combination of 1 hidden layer, with 18 neurons was selected as the best result, since it keeps a lower number of parameters, even though it is not significantly better than other combinations.

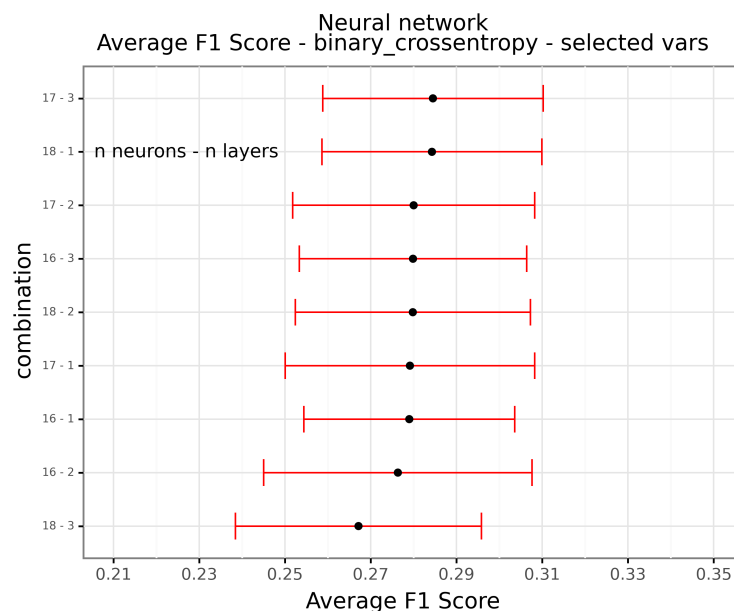


FIGURE C.23: Neural network results. UAI.

For Universidad de Talca, once again, it was observed that a low number of layers (less than 20) produces results with less variance. After discarding architectures with a large number of layers, we proceed to search for the best set of variables. The variables selected were: mat, pps, admission, nem,

lang, and preference. As in UAI, the combination of 1 hidden layer, with 18 neurons was selected as the best result, even though it is not significantly better than the other combinations.

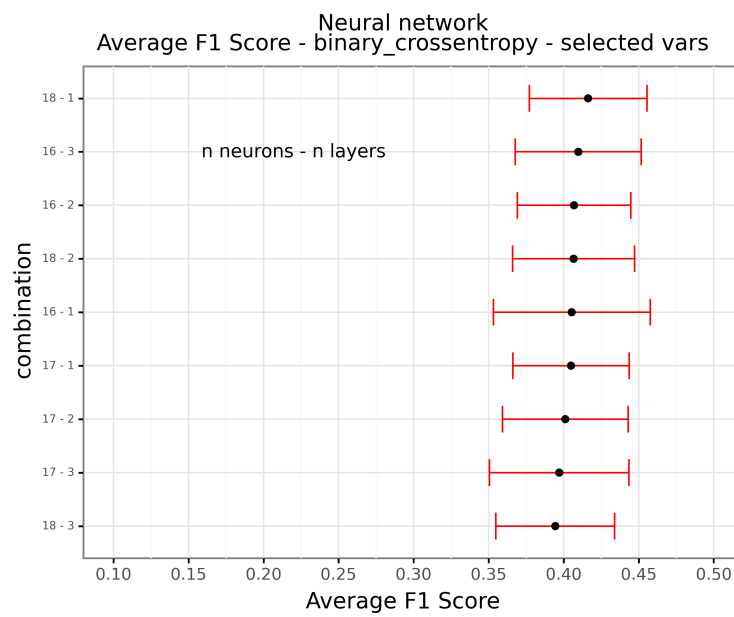


FIGURE C.24: Neural network results. U de Talca.