Un journal de asfae chile

mbr Multidisciplinary Business Review

AN APPLICATION OF MACHINE LEARNING IN PUBLIC POLICY. EARLY WARNING PREDICTION OF SCHOOL DROPOUT IN THE CHILEAN PUBLIC EDUCATION SYSTEM

UNA APLICACIÓN DE APRENDIZAJE AUTOMÁTICO (*MACHINE LEARNING*) EN POLÍTICAS PÚBLICAS. PREDICCIÓN DE ALERTA TEMPRANA DE DESERCIÓN ESCOLAR EN EL SISTEMA DE EDUCACIÓN PÚBLICA DE CHILE

Jerome Smith Uldall[a] • Cristián Gutiérrez Rojas[b]

Classification: Empirical paper –case study
Received: December 31, 2020 / Revised: March 16, 2021; October 1, 2021 / Accepted: November 22, 2021

Abstract

School dropout is a serious problem worldwide, and contributes to a great deal of poverty and misery. People who have not finished school obviously suffer the consequences, but these extend to all of society since they become a burden due to lack of education and skills for the workplace. Much like poverty, school dropout is complex and multidimensional. Hence, early warning systems that predict which children are at risk of dropping out of school are of the utmost importance, and furthermore, the interventions to rescue these children must be bespoke, i.e., tailored to the specific situation of each child. Much work has been done using traditional methods such as attendance thresholds and logistic regression. However, school dropout prediction by means of applying machine learning is relatively new. In addition, an application that has worked in one country does not necessarily work in another, since the available data sets are different. Therefore, the following question arises: does machine learning enable a more accurate early warning of school dropout specifically in Chile? In this paper we answer this question, testing and comparing machine learning predictive models with a traditional logistic regression, using public databases from the Chilean Ministry of Education. In addition, we offer some practical recommendations for other researchers and policy makers who endeavour to implement practical working early warning systems for school dropout.

Keywords: Machine learning, school dropout, predictive model, neural network, decision tree.

Resumen

La deserción escolar es un grave problema social mundial que contribuye a mucha pobreza y sufrimiento. Las personas que no han terminado la escuela obviamente sufren las consecuencias directas, pero además extienden este problema al resto de la sociedad, ya que se convierten en una carga social debido a su falta de educación y capacitación laboral. Al igual que la pobreza, la deserción escolar es compleja y multidimensional. Por tanto, los sistemas de alerta temprana que predicen los niños específicos que están en riesgo de desertar son de suma importancia. Más aún, las intervenciones para rescatar a estos estudiantes tienen que hacerse a la medida de la situación social y emocional específica de cada niño. Se ha hecho bastante trabajo en este sentido, usando métodos tradicionales como umbrales de asistencia y regresión logística. Sin embargo, la predicción de la deserción escolar aplicando métodos de *machine learning* es relativamente nuevo. Además, una aplicación que ha funcionado en un país no necesariamente funciona

[a] Universidad Católica Silva Henríquez, Santiago, Chile. ORCID: 0000-0003-0967-1663. Email: jsmith@ucsh.cl
[b] Universidad Católica Silva Henríquez, Santiago, Chile. Email: cgutierrez@ucsh.cl

en otro, porque los datos disponibles en cada país son distintos. Por tanto, surge la pregunta: ¿Puede el *machine learning* brindar sistemas de alerta temprana de deserción escolar más certeros específicamente en Chile? En este artículo, respondemos esta pregunta, probando y comparando modelos predictivos de *machine learning* con una regresión logística tradicional, usando la base de datos pública del Ministerio de Educación de Chile. Además, ofrecemos algunas recomendaciones prácticas para investigadores y gestores de políticas públicas que deseen implementar sistemas reales de alerta temprana de deserción escolar.

Palabras clave: aprendizaje automático, deserción escolar, modelo predictivo, red neuronal, árbol de decisiones.

Introduction and Literature Review

School dropout is a serious problem worldwide, particularly in Latin America where it reaches rates of 30% (Espíndola & León, 2002), and contributes to a great deal of poverty and misery, according to many studies (Adelman & Székely, 2017). The loss of income suffered by those who drop out of school ranges between 20% and 40%, according to Espíndola and León (2002). Moreover, school dropout is greater amongst children who are from poor backgrounds, thus exacerbating inequality in a vicious cycle. Clearly, much can be done to improve the quality of life and overcome inequality by tackling school dropout.

Moreover, even though many studies point to poverty and inequality as the underlying root cause of school dropout, the reasons for dropping out as stated by young people are diverse: work, teenage pregnancy, lack of interest, etc. (Melis et al., 2005). In summary, the causes and nature of school dropout are multidimensional. This is the reason why many government agencies, in particular the Chilean Ministry of Education (Mineduc), have adopted a personalised approach to school dropout, that is, predicting and identifying the children that are most at risk of dropping out of school, and then intervening appropriately with bespoke measures targeted at the particular situation of each child. This intervention is supported and carried out by a dual team made up of a psychologist and a social worker ("dupla psicosocial"), one assigned to each school (Mineduc, 2014). Therefore, it is clear that an early warning predictive model that identifies, with a high degree of accuracy, the children that are most at risk of dropping out of school, would be a very valuable tool in supporting these bespoke interventions.

The importance of increasing both measures of accuracy: specificity and sensitivity, is reflected in the social costs of their counterparts: false negatives and false positives. The former represents the children who are not detected by the early warning system and hence drop out without having had the opportunity to receive help. The latter begets waste of scarce resources that could be better used elsewhere, in particular, addressing the root causes of school dropout.

Early warning systems of school dropout have usually applied traditional methods, such as attendance and marks/grades thresholds (assuming that children below a certain threshold are at greater risk of dropping out) and logistic regression. Examples are the SIAT early warning system of Mexico (Secretaría de Educación Pública, 2011), the SAT of the Municipality of Peñalolén, Chile (Municipalidad de Peñalolén, 2012), and the prediction of university dropout in Ecuador (Alban & Sánchez, 2018).

Artificial Neural Networks (ANN) have been used increasingly in econometric research and forecasting since at least the turn of the century. In particular, in 2000 Moshiri & Cameron demonstrated that ANN performed at least as well as traditional methods in forecasting inflation.

In Varian's 2014 paper, "Big Data: New Tricks for Econometrics", the author proposed a challenge to economists: supersede traditional methods using big data and machine learning in econometrics. These days it is indisputable that data science (essentially statistical methods applied to big data) has acquired great relevance in solving complex problems in numerous practical fields (The Economist Intelligence Unit, 2015), and it is expected that its penetration will be even greater in the future. However, handling massive data sets of millions of observations in relational databases, coupled with sophisticated machine learning techniques, can be challenging for economists trained in traditional econometric methods with comparatively small data sets that fit on a spreadsheet (Einav & Levin, 2014; Varian, 2014). Nevertheless, the rewards in terms of greater accuracy and predictive power, and the possibility of modelling non-linear phenomena, make applying these new techniques worthwhile (Mduma & Neema, 2019).

Indeed, Mullainathan and Spiess in their 2017 paper "Machine Learning: An Applied Econometric Approach" provide a valuable theoretical foundation for applying machine learning in forecasting and public policy and highlight many exciting new applications, for example satellite imagery to predict poverty and economic output.

Machine learning for predicting school dropout has been applied successfully in Brazil (Barros et al., 2019;

Martinho et al., 2013) and South Korea (Lee & Chung, 2019). However, at the time in which our work was carried out (2015), to the best of our knowledge it had not yet been applied to school dropout in Chile, with the data available in that country.

Therefore, our question is: does machine learning enable more accurate early warning of school dropout in Chile? Our hypothesis is that it does, significantly. In this paper we "pick up Hal Varian's gauntlet" and present an application of large data sets with machine learning techniques in public policy: an early warning predictive model of school dropout in the Chilean public education system.

This paper consists of the following sections in addition to this introduction. The "method" section describes the database used and the methodology applied for its treatment in addition to the preliminary data analysis, including scatterplots for each selected variable and correlation tables. Moreover, we briefly describe the binary choice econometric and machine learning models used. In the following section the results are presented, in terms of the accuracy, specificity and sensitivity of each of the models. Finally, in the "conclusions" section we discuss some conclusions and recommendations for the design of a future predictive model for school dropout in Chile, which we hope will stimulate further work and research.

Method

This paper complies with the principles of reproducible research, and therefore all the data sources are clearly specified. The source code used, both SQL and R, may be requested from the authors by email.

Data Sources

The data set we used was a freely available public database provided by the Chilean Ministry of Education on its website. www.mineduc.cl

Table 1. Dropout Rate

Year	Registered Students	Should be Registered	Dropout	Dropout Rate
2011	2 999 885	3 098 348	98 463	3.18%
2012	2 933 483	3 052 797	119 314	3.91%
2013	2 901 688	2 999 209	97 521	3.25%

Source: Government of Chile, Ministry of Education, 2018. http://datosabiertos.mineduc.cl/

At the time we carried out this study, we downloaded the data from the following link: http://centroestudios.mineduc.cl/

At the time we submitted this paper for publication, the Ministry's data download link had changed to: http://datosabiertos.mineduc.cl/

The data consisted of student registration information, attendance records, school marks/grades and teacher data from a four-year period, from 2011 to 2014, of all schools excluding private ones.

We also downloaded socioeconomic data by commune (the smallest Chilean geographical administrative subdivision) from the Chilean national institute of statistics: http://www.ine.cl/

We studied the cohort of students registered at school in the years 2011-2013, and followed up each student's process during this three-year period. The 2014 registration data was used only to calculate the students that dropped out in 2013.

Table 2. Prevalence Rate

Year	Population of School Age (6-21)	Not in School	Dropout Prevalence
2011	3 509 668	318 614	9.08%
2012	3 458 065	300 088	8.68%
2013	3 417 535	305 171	8.93%

Source: Government of Chile, Ministry of Education, 2018. http://datosabiertos.mineduc.cl/

Almost all of the downloaded files were compressed with Winrar. To uncompress them, the Winrar software must be downloaded and installed. The link is: https://www.win-rar.com

The downloaded files are displayed in Table 3.

Table 3. Downloaded Files

N°	File	Source
1	2011 Matrícula por estudiante.rar	http://centroestudios.mineduc.cl/
2	2012 Matrícula por estudiante.rar	http://centroestudios.mineduc.cl/
3	2013 Matrícula por estudiante.rar	http://centroestudios.mineduc.cl/
4	2014 Matrícula por estudiante.rar	http://centroestudios.mineduc.cl/
5	Docentes 2011.rar	http://centroestudios.mineduc.cl/
6	Docentes 2012.rar	http://centroestudios.mineduc.cl/
7	Docentes 2013.rar	http://centroestudios.mineduc.cl/
8	Docentes 2014.rar	http://centroestudios.mineduc.cl/
9	Rendimiento por estudiante 2011.rar	http://centroestudios.mineduc.cl/
10	Rendimiento por estudiante 2012.rar	http://centroestudios.mineduc.cl/
11	Rendimiento por estudiante 2013.rar	http://centroestudios.mineduc.cl/

(*Continued*)

N°	File	Source
12	Rendimiento por estudiante 2014.rar	http://centroestudios.mineduc.cl/
13	Comunas Socioeconomicos.csv	Datos http://www.ine.cl/

Source: Government of Chile, Ministry of Education, 2018. http://datosabiertos.mineduc.cl/. Instituto Nacional de Estadísticas, 2018. http://www.ine.cl/

The files were all in csv format, one for each year, registered student and school year result. The precise specification of all the uncompressed files downloaded is in Table 4.

Table 4. Uncompressed Source Data Files

N°	File	Description
1	20140812_matricula_unica_2011_20110430_PUBL.csv	Registered students in year 2011
2	20140812_matricula_unica_2012_20120430_PUBL.csv	Registered students in year 2012
3	20140808_matricula_unica_2013_20130430_PUBL.csv	Registered students in year 2013
4	20140924_Matrícula_única_2014_20140430_PUBL.csv	Registered students in year 2014
5	20130301_Rendimiento_2011_20120416_PUBL.csv	Grades and attendance in year 2011
6	20130515_Rendimiento_2012_20130430_PUBL.csv	Grades and attendance in year 2012
7	20140227_Rendimiento_2013_20140206_PUBL.csv	Grades and attendance in year 2013
8	20150225_Rendimiento_2014_20150218_PUBL.csv	Grades and attendance in year 2014
9	docentes 2011.csv	Teacher data in year 2011
10	docentes 2012.csv	Teacher data in year 2012
11	20130904_Docentes_2013_20130709_PUBL.csv	Teacher data in year 2013
12	20140819_Docentes_2014_20140704_PUBL.csv	Teacher data in year 2014
13	Comunas Datos Socioeconomicos.csv	Socioeconomic data by commune

Source: Government of Chile, Ministry of Education, 2018. http://datosabiertos.mineduc.cl/

Software Used: SQL Server and R

To process and clean the data we used SQL language with Microsoft SQL Server Evaluation Edition 2017, which can be freely downloaded from Microsoft Corporation's website: www.microsoft.com/sql

The SQL source code also runs in earlier versions of SQL Server.

To perform the statistical analysis and modelling, we used Microsoft R client, the free version of R supplied by Microsoft that includes the RevoscaleR package that is capable of handling big data. It can be downloaded from: https://docs.microsoft.com/en-us/machine-learning-server/r-client/install-on-windows

Definition of Variables

Dropout is represented by a binary variable, in which 1 represents dropping out and 0 not dropping out.

Our definition of dropping out is based on the following rules:

If a student is registered in year t and not registered in year t+1, dropout occurs.

In addition to rule (1), if the student is registered in adult education in year t+1, dropout occurs.

If the student in year t is registered in the final year of secondary school and is flagged with a pass (defined by passing marks and attendance), then dropout does not occur.

We performed a preliminary exploration of the data and based on a combination of literature reports and common sense, we selected a set of variables that in our view may be predictors of dropout, or at least interesting to analyse.

Table 5 describes all the candidate predictive variables, with their respective definitions.

Table 5. Candidate Predictive Variables

N°	Variable	Description
1	Attendance	School attendance: 0% - 100%.
2	SchoolMarks	School marks or grades: 0% - 100%, converted from Chilean scale of 1 - 7 for international readability.
3	TotalSchoolChanges	Total number of times that the student has changed schools.
4	YearlyAverageClassChanges	Average number of times that the student has changed class in each year.
5	SchoolOrd	School ordered by probability of dropout.
6	EducationLevelOrd	School Level ordered by probability of dropout.
7	SchoolYear	School year starting at year 1 for grade 1 of Primary School, up to year 12 for last year (4th grade) of Secondary School.
8	EducationTypeOrd	Education Type ordered by probability of dropout.
9	TotalSchoolYearRepeats	Total number of times that the student has repeated the same level, usually due to failing that level.
10	StudentsPerClass	Number of students per class.
11	TeacherContractHours	Number of hours of each teacher according to the work contract.
12	TeacherYearsService	Number of years of service of teacher.
13	TeacherYearsInSchool	Number of years teacher has taught at the student's school.

(*Continued*)

N°	Variable	Description
14	StudentsPerSchool	Number of children per school.
15	Age	Age of student.
16	AgeDifference_wr_LevelGrade	Age difference with respect to the standard age expected for the level and grade.
17	Gender	Gender: M=male, F=female.
18	StudentCommune	Commune of residence of student.
19	SchoolCommune	Commune of school.
20	MeanIncome_StudentCommune	Mean income of commune of residence of student.
21	MeanIncome_SchoolCommune	Mean income of commune of school.

(*Continued*)

Source: Government of Chile, Ministry of Education, 2018. http://datosabiertos.mineduc.cl/

There are 6 synthetic ordinal variables: SchoolOrd, EducationLevelOrd, EducationTypeOrd, GenderOrd, StudentCommuneOrd and SchoolCommuneOrd, that we derived from their corresponding categorical variables (SchoolID, etc.) by ordering their categorical values by the dropout probability. This is to avoid creating a very large number of dummy variables (there are over 9000 schools and 346 communes). Synthetic ordinal variables can be used for prediction purposes, although clearly, they have no meaningful interpretation because any categorical variable can be made to correlate with a probability by ordering the values by probability. For this reason, we have omitted the scatterplots of the synthetic ordinal variables.

Table 6. EducationLevelOrd Mapping

Education Level ID	Name	Education Level Ord	Dropout Probability
2	Primary Education	1	5.96%
1	Nursery / Pre School	2	10.45%
4	Secondary Education	3	14.99%
6	Secondary Technical Education	4	15.58%

Source: Government of Chile, Ministry of Education, 2018. http://datosabiertos.mineduc.cl/

The mapping of original categorical values for EducationLevelOrd, EducationTypeOrd and GenderOrd are shown on the tables 6, 7 and 8. (For space reasons we have omitted the mappings for SchoolOrd, StudentCommuneOrd and SchoolCommuneOrd).

Table 7. EducationTypeOrd Mapping

Education Type ID	Name	Education Type Ord	Dropout Probability
110	Primary Education	1	5.96%
10	Nursery / Pre School	2	7.59%
214	Special Education - Visual Disability	3	8.70%
299	Education Integration Programme (PIE in Spanish)	4	10.00%
610	Secondary Technical Education	5	13.46%
410	Secondary Technical Education - Commercial	6	14.79%
310	Secondary Education	7	14.99%
510	Secondary Technical Education - Industrial	8	16.59%
810	Secondary Technical Education - Maritime	9	16.77%
710	Secondary Technical Education - Agricultural	10	19.88%
216	Special Education - Autism	11	28.57%
910	Secondary Technical Education - Artistic	12	33.33%
212	Special Education - Intellectual Disability	13	45.65%
211	Special Education - Auditive Disability	14	85.71%
215	Special Education - Motor Disability	15	100.00%

Source: Government of Chile, Ministry of Education, 2018. http://datosabiertos.mineduc.cl/

Table 8. GenderOrd Mapping

Gender	Gender Ord	Dropout Probability
F	1	7.45%
M	2	9.63%

Source: Government of Chile, Ministry of Education, 2018. http://datosabiertos.mineduc.cl/, and own elaboration.

Data Cleaning

The aim of the data cleaning phase was to produce a tidy table of observations. Each observation consists of a student registered in at least one of the years of the three-year period, and the associated result of dropping out or not. In addition, each observation has the associated predictive variables, such as attendance, marks, type of education, etc.

We did not consider the death rate of students during the three-year period, due to this information not being available in the data, but we assume that at school age this factor is negligible and does not affect our results.

The complete list of columns (including variables) on the observations table is shown in Table 9.

Table 9. Observations Table

N°	Column	Role
1	StudentID	Primary key
2	Dropout	Independent variable
3	Attendance	Predictive variable
4	SchoolMarks	Predictive variable
5	TotalSchoolChanges	Predictive variable
6	YearlyAverageClassChanges	Predictive variable
7	SchoolOrd	Predictive variable
8	School_Label	Variable label, mainly for plots
9	EducationLevelOrd	Predictive variable
10	EducationLevel_Label	Variable label, mainly for plots
11	SchoolYear	Predictive variable
12	EducationTypeOrd	Predictive variable
13	EducationType_Label	Variable label, mainly for plots
14	TotalSchoolYearRepeats	Predictive variable
15	StudentsPerClass	Predictive variable
16	TeacherContractHours	Predictive variable
17	TeacherYearsService	Predictive variable
18	TeacherYearsInSchool	Predictive variable
19	StudentsPerSchool	Predictive variable
20	Age	Predictive variable
21	AgeDifference_wr_ LevelGrade	Predictive variable
22	GenderOrd	Predictive variable
23	Gender_Label	Variable label, mainly for plots
24	StudentCommuneOrd	Predictive variable
25	StudentCommune_Label	Variable label, mainly for plots
26	SchoolCommuneOrd	Predictive variable
27	SchoolCommune_Label	Variable label, mainly for plots
28	MeanIncome_ StudentCommune	Predictive variable
29	MeanIncome_ SchoolCommune	Predictive variable

Source: Government of Chile, Ministry of Education, 2018. http://datosabiertos.mineduc.cl/ and own elaboration.

The table consisted of a pool of 3 399 147 observations and 22 variables, including the dependent variable, Dropout.

The entire data cleaning process consisted of 69 steps. For space reasons, the precise steps are not included here. If the reader wishes to attain exactly the same results, they may write to the authors and we will be happy to provide the details and the source code.

In order to verify the predictive power of the models, the two final steps of the above process consisted of randomly dividing the complete observations table into two subsets: 80% for training and 20% for testing.

Exploratory Data Analysis

In order to acquire some intuitive insight regarding the data and the relationships between the variables, we performed an exploratory data analysis.

Firstly, with the aim of identifying possible linear relationships, we generated a table of the correlation of each predictive variable with the dropout variable, and the correlations between the predictive variables.

Secondly, we created histograms and scatterplots of all the meaningful variables.

Algorithms Used

We tested and compared a traditional econometric algorithm: logistic regression, with three machine learning ones: decision trees, random forests and neural networks. The reason for this selection is as follows. Logistic regression is a traditional econometric binary model technique that is used frequently and taught in many undergraduate econometrics courses (Dougherty, 2011). In addition, decision trees, random forests and neural networks are widely cited in the academic literature (Moshiri & Cameron, 2000; Mullainathan & Spiess, 2017).

Below is a brief description of each algorithm used. The precise mathematical details are beyond the scope of this paper, but the interested reader may study them in further depth (James et al., 2017; Poggio & Smale, 2003).

If Y is the binary dependent variable, with possible values {0, 1}, then the probability of Y being equal to 1 (the event occurring) is given by the logit probability function:

$$P(Y=1)=\frac{1}{1+e^{-Z}}\ldots \quad (3)$$

where Z is a linear function of the predictive variables:

$$Z=\beta_0+\beta_1 X_1+\ldots+\beta_K X_k\ldots \quad (4)$$

The Figure of p against Z is a sigmoid with asymptotes p = 0 and p = 1.

To fit the model, the algorithm uses maximum likelihood estimation to find estimators b_0, b_1, … b_k for the coefficients β_0, β_1, … β_k.

For further information, see Kleinbaum et al. (2010).

Decision Trees

The algorithm first ranks the predictive variables by their correlation with the dependent variable. Starting with the variable with the highest correlation, it divides the data set into subsets according to the frequency of the depend-

ent variable. For each variable it further subdivides the data set. Each node of the tree corresponds to a subset. The final layer of the tree (the leaf nodes) is the set of subsets obtained by applying all the variables, and each one has the associated probability of the dependent variable.

Random Forest

Random forests work by constructing a large number of decision trees and then voting for the best tree.

Neural Networks

Neural networks are inspired by the structure of the human brain, in which neurons are connected to each other by dendrites and axons.

The input layer of neurons is mapped 1-1 to the predictive variables X_1, … X_k. Each input neuron is connected to every neuron of the next layer, the first hidden layer. The intensity of the signal from input neuron i to hidden layer neuron j is given by the weight w_{ij}. There may be several hidden layers. The last hidden layer is connected to the output layer, which has one neuron for each dependent variable. Training occurs by comparing the output value with the real value and then adjusting the weights w_{ij}.

Selection of Variables

As a first step, we excluded variables for which it is a reasonable assumption that any relationship they have with dropout must be circumstantial and hence spurious, since these variables are most likely correlated with the variables that are more fundamentally correlated with dropout. One example is the school. Some schools have a much higher dropout rate than others, which makes the school a good predictor of dropout, *but only in the present, for the particular data set we used*. This variance between schools is most likely due to circumstantial differences in the running of each school which would affect the underlying fundamental variables. However, the performance of each school will very likely change over time, and therefore in the future the predictive power of the school will deteriorate over time. For the same reason we also eliminated Student Commune and School Commune, which will most likely change under public policy that aims to reduce school dropout.

For the selection of all the remaining variables, we performed the following algorithm written in R, in two phases.

In the first phase, we took each predictive variable in turn and trained a simple decision tree model of dropout with only this predictive variable. For each *univariate model* we calculated the accuracy, specificity (true negative rate) and sensitivity (true positive rate). Since the specificities were all very similar (above 99%), we ranked the variables in descending order by the sensitivity of their corresponding univariate models. See Table 12.

In the second phase, we started with the univariate model of the highest-ranking variable, and added the remaining variables one by one, in the ranking order of phase (1), registering the accuracy, specificity and sensitivity of each corresponding *cumulative model*. We then calculated the incremental sensitivity, **Δ***Sensitivity*, contributed by each variable. See Table 13. We then eliminated the variables that did not add significant sensitivity to the cumulative model (the variables highlighted in grey on Table 13).

The remaining variables in the table are those selected for the final models, described in the following section.

Training, Testing and Comparison of Models

With the selected variables we applied the aforementioned four algorithms to train and test the models. We trained the models with the training subset of 80% of the observations and tested them with the 20% testing subset.

To compare the models, we generated the individual confusion matrices and calculated their respective accuracies, true negative rates (specificities) and true positive rates (sensitivities).

Results

Exploratory Data Analysis

Correlation between Variables and Dropout

The Table 10 presents the correlation coefficients between the candidate predictive variables and dropout. The absolute values greater than 0.1 are highlighted in grey.

Table 10. Correlation Coefficients

Variable	Coefficient
Dropout	1
TotalSchoolYearRepeats	0.346
AgeDifference_wr_Level Grade	0.315
SchoolOrd	0.248
Age	0.223
SchoolYear	0.173
TotalSchoolChanges	0.158
EducationTypeOrd	0.149
Education LevelOrd	0.146

(*Continued*)

Variable	Coefficient
SchoolCommuneOrd	0.058
StudentCommuneOrd	0.052
GenderOrd	0.038
TeacherYearsService	0.021
Teacher YearsInSchool	0.014
MeanIncome_SchoolCommune	0.002
YearlyAverageClassChanges	0.001
MeanIncome_StudentCommune	-0.000
StudentsPerSchool	-0.027
StudentsPerClass	-0.028
TeacherContractHours	-0.033
Attendance	-0.226
School Marks	-0.331

Source: Government of Chile, Ministry of Education, 2018. http://datosabiertos.mineduc.cl/ and own elaboration.

The Figure 1 clearly illustrates the above correlation coefficients.

In Table 10 and Figure 1 two distinct groups are clearly apparent, above and below the cut-off value of +/- 0.1. According to this criterion the variables with the highest linear correlation with dropout are: Total School Year Repeats, School Marks, Age Difference with respect to Level/Grade, School, Attendance, Age, School Year, Total School Changes, Education Type and Education Level.

Variables, Histograms and Scatterplots

In this section we present the interpretations of the histograms and scatterplots of only those variables with absolute value of correlation coefficient greater than 0.1. It is important to emphasise that the reason for this selection is solely for expository purposes, and that at this stage we were not yet selecting variables for the predictive model.

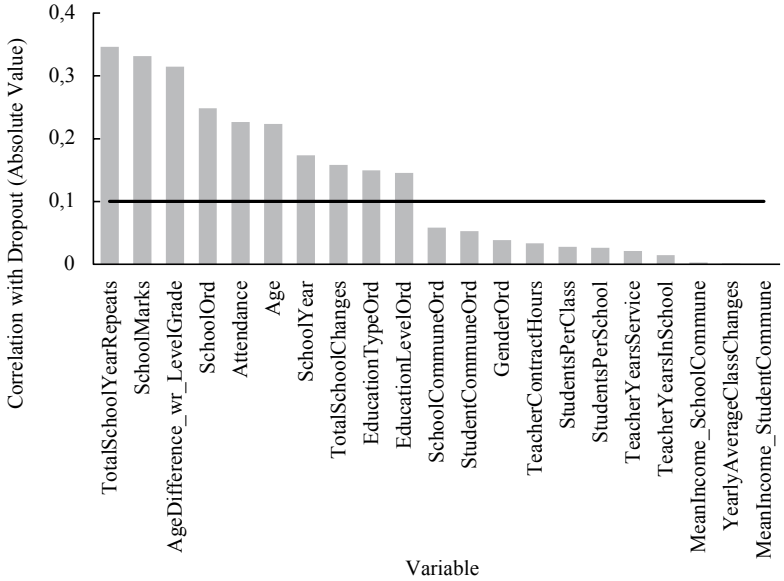
Total School Year Repeats

The high correlation coefficient of this variable appears to indicate a significantly strong relationship with dropout, which is consistent with intuition. This quantitative variable represents the total number of times that the student has repeated the same level and grade, usually due to failing the grade. Four values are observed in the data: 0 (the majority of cases), 1, 2 and 3. (Figure 2).

School Marks/Grades

One indicator of academic performance is school marks (grades). Clearly there is a negative relationship between this variable and dropout, confirming the idea that dropout is associated with academic failure. For international readability this variable is expressed in the range of 0% - 100%, converted from the Chilean scale of 1-7, where a passing mark is 50% (4 on the Chilean scale). (Figure 3).

Figure 1. Correlation Coefficients (absolute value) of Candidate Variables with Dropout



Source: own elaboration.

Figure 2. Total School Year Repeats Scatterplot and Histogram

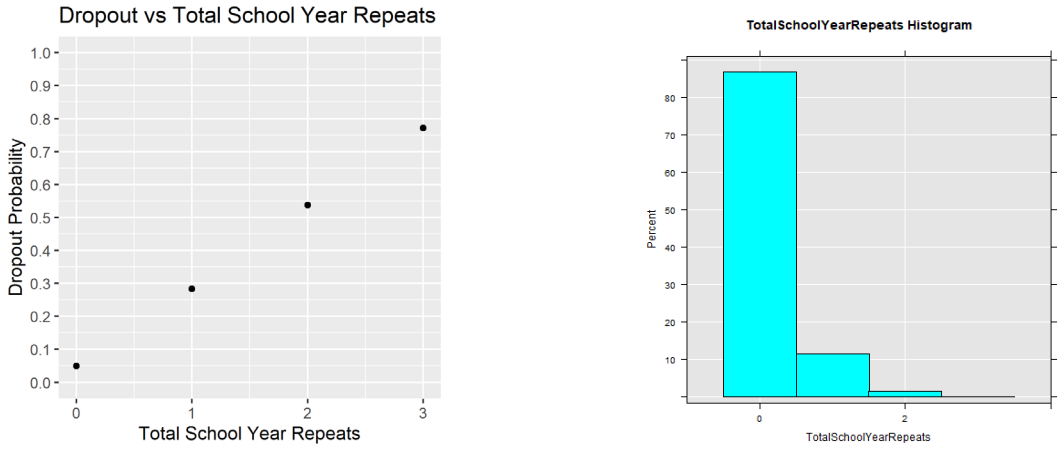
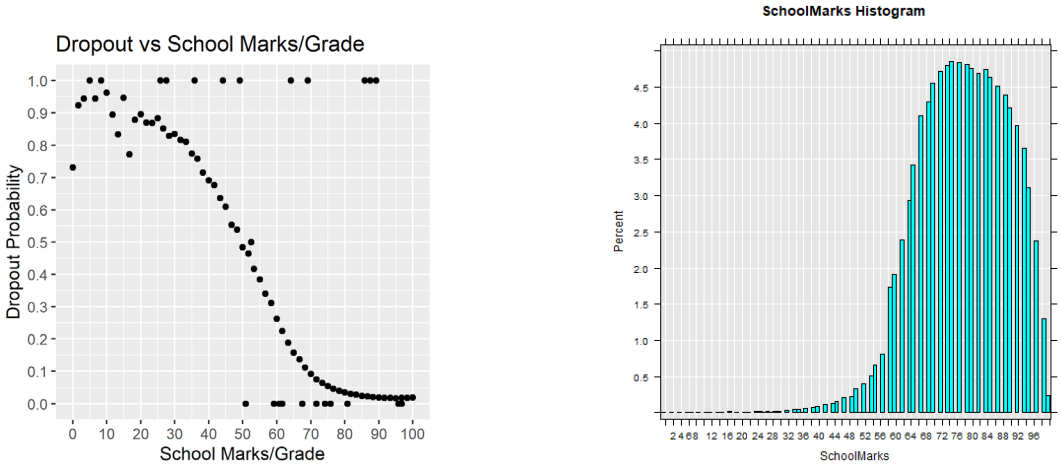


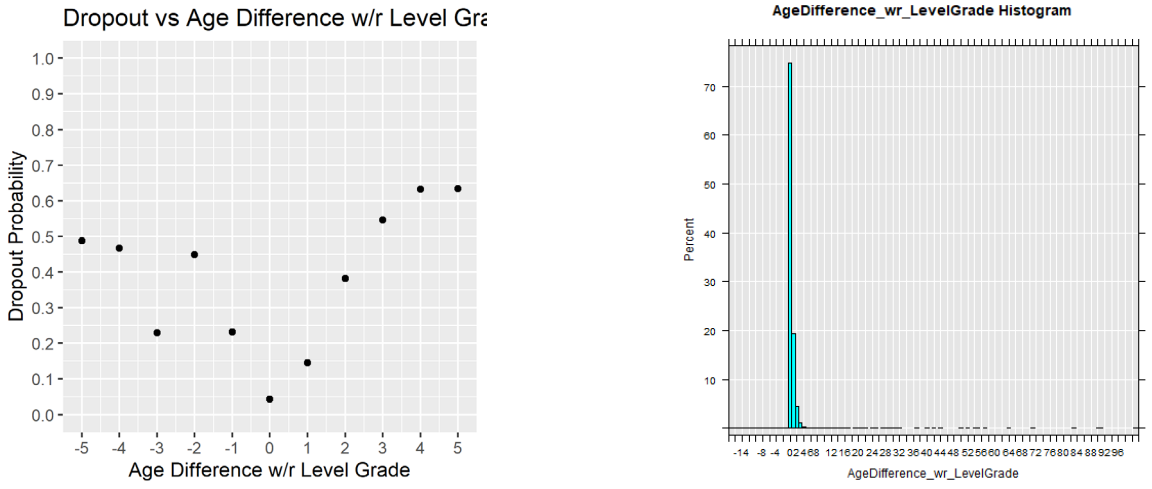
Figure 3. School Marks Scatterplot and Histogram



Age Difference with respect to Education Level

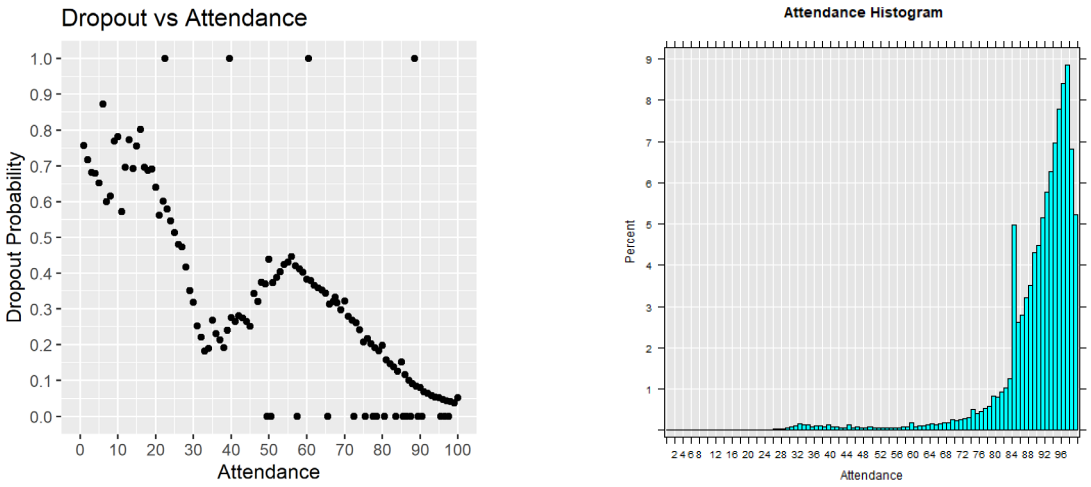
The scatterplot is roughly symmetrical, with a minimum at zero, suggesting that both negative and positive differences between the age and the grade level are positively associated with dropout. This variable indicates the age difference with respect to the standard age expected for the level and grade according to the Chilean educational system, composed of 12 levels; the first eight correspond to primary education, while the last four correspond to secondary education. The range of differences displayed is between -5 and 5 years (Figure 4).

Figure 4. Age Difference with respect to Scatterplot and Histogram



Multidiscip. Bus. Rev. | *Vol. 15, N° 1, 2022, pp. XX-XX, ISSN 0718-400X*
DOI: *https://doi.org/*

Figure 5. Attendance Scatterplot and Histogram



Attendance

This is one of the most frequently cited variables in the academic literature regarding school dropout. The scatterplot displays a non-trivial relationship, with three segments. First, from 0 to around 35 percent a negative relationship, where higher attendance implies lesser dropout. Second, from 35 to 55 percent a positive relationship, and finally from 55 to 100 percent a negative relationship again. This means that attendance does not have a linear relationship with dropout. The histogram shows the normal behaviour for this kind of variable (left-skewed). (Figure 5)

Age

We can observe a peak in dropout probability during the teenage years, corresponding to secondary education. This has important implications for public policy since dropout prevention measures can be focused on this age group (Figure 6).

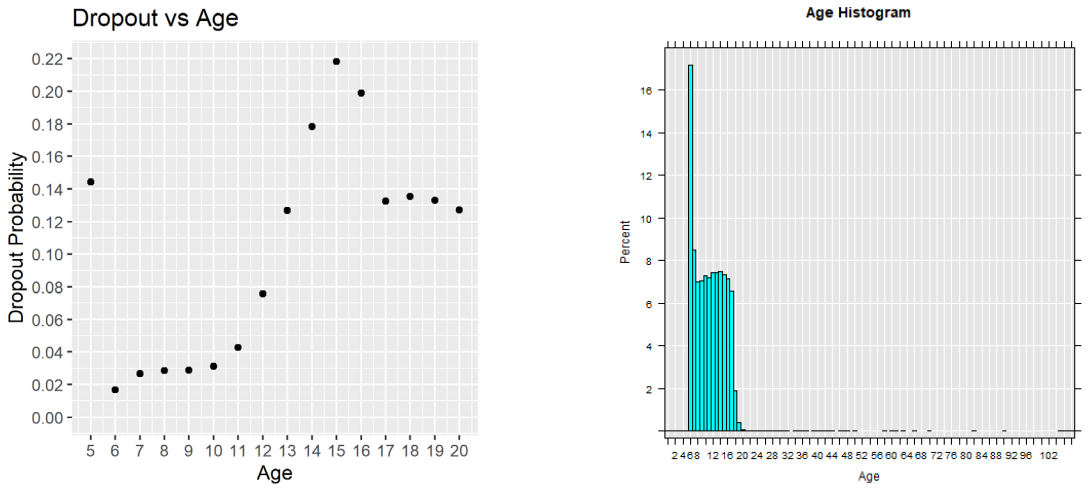
School Year

The Chilean educational system comprises 12 years of compulsory education: 8 years of primary school and 4 years of secondary school. The school year is simply the sequential number of the year of each level, starting with year 1 for level 1 of primary school and ending at year 12 for level 4 of secondary school. The most important year in percentile terms is year 1.

Between years 1 and 9 the dropout rate rises, and then falls from year 9 to 12.

We observe that the greatest dropout occurs at year 9, one year after the transition from primary to second-

Figure 6. Age Scatterplot and Histogram



ary school (from year 8 to 9). The above results could be related to the social importance attached to completing an educational cycle, especially those conducive to certification, such as the completion of primary education. Moreover, it is important to consider that a significant number of schools do not teach secondary education and, therefore compel students who finish primary school to seek another school for their secondary education, thus increasing the probability of dropping out in these cases. (Mineduc, 2014). (Figure 7).

Total School Changes

This is the total number of times that the student changes schools. Our hypothesis is that a child who changes school often undergoes disruption in his/her life which may be a factor encouraging dropout, or at least is correlated with other factors such as an unstable family life, changes in socioeconomic status, etc, which in themselves are possible causes of dropout. In effect, the scatterplot shows that the probability of dropout increases with this variable (Figure 8).

Students per Class

This is the number of students in each class. The histogram shows that most classes have 20 – 50 students. In this range the dropout probability appears to decrease with class size, which may be counterintuitive given educational policy that recommends limiting the number of students per class (Figure 9).

Figure 7. School Year Scatterplot and Histogram

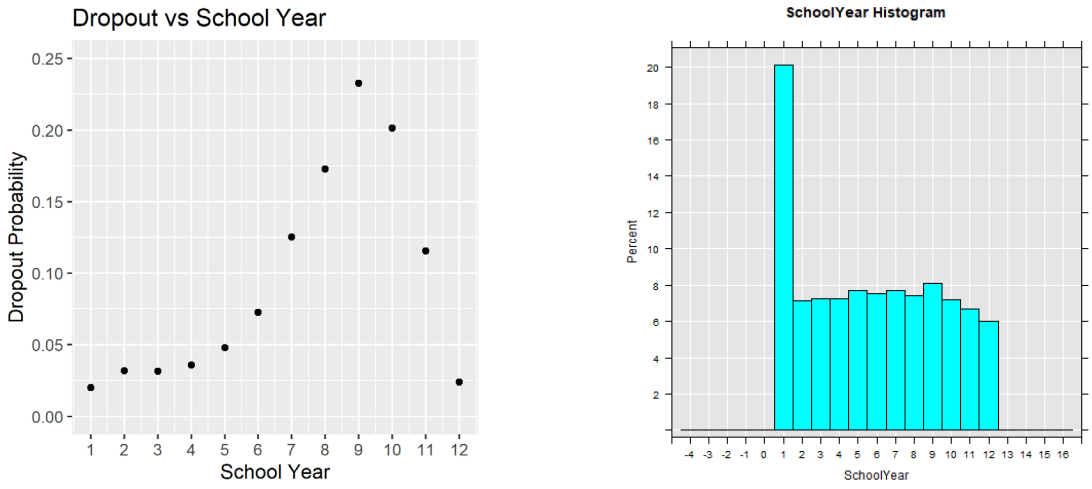
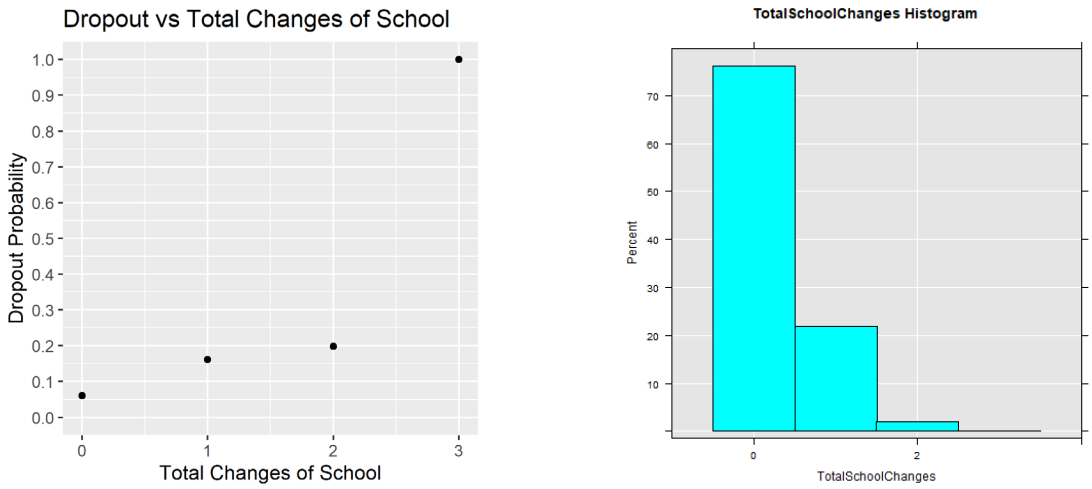
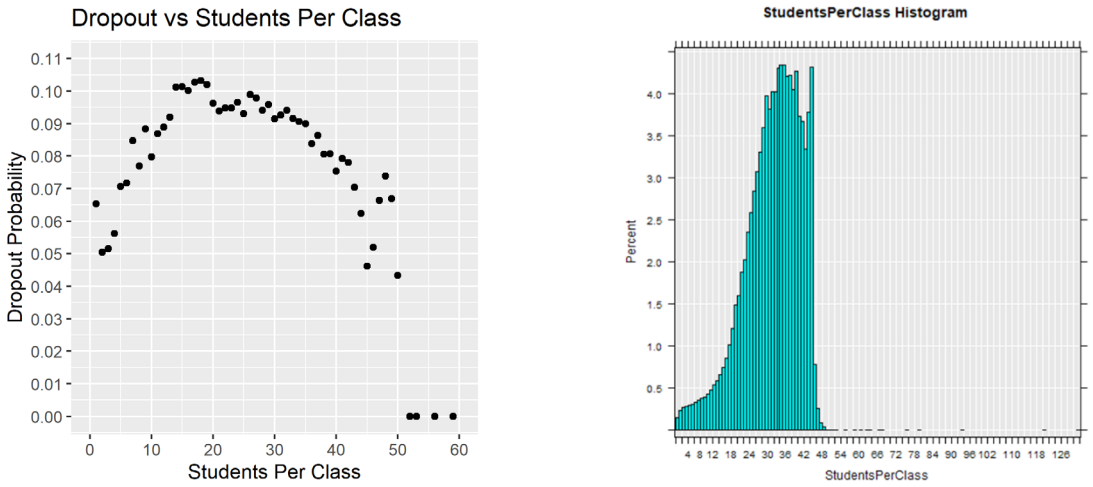


Figure 8. Total School Changes Scatterplot and Histogram



Multidiscip. Bus. Rev. | *Vol. 15, N° 1, 2022, pp. XX-XX, ISSN 0718-400X*
DOI: *https://doi.org/*

Figure 9. Students per Class Scatterplot and Histogram



Students per School

Similar to the above is the number of students per school. The scatterplot shows no apparent trend (Figure 10).

Correlation Table

The correlation matrix of all for all variables, using the entire data set, is shown in Table 11.

Selected Variables

Phase 1 of the variable selection algorithm produced the Table 12 of individual sensitivities.

Phase 2 of the variable selection algorithm produced the Table 13 of cumulative sensitivities.

The rightmost column shows the increase in sensitivity, ΔSensitivity. Zero and negative values, and the corresponding variables, are highlighted in grey. These variables do not contribute any sensitivity to the model; hence it makes sense to eliminate them.

Therefore, the variables selected for the final models are the following:

TotalSchoolYearRepeats
SchoolMarks
AgeDifference_wr_LevelGrade
Attendance
TotalSchoolChanges
EducationLevelOrd
SchoolYear
StudentsPerClass
StudentsPerSchool
Age

Figure 10. Students per School Scatterplot and Histogram

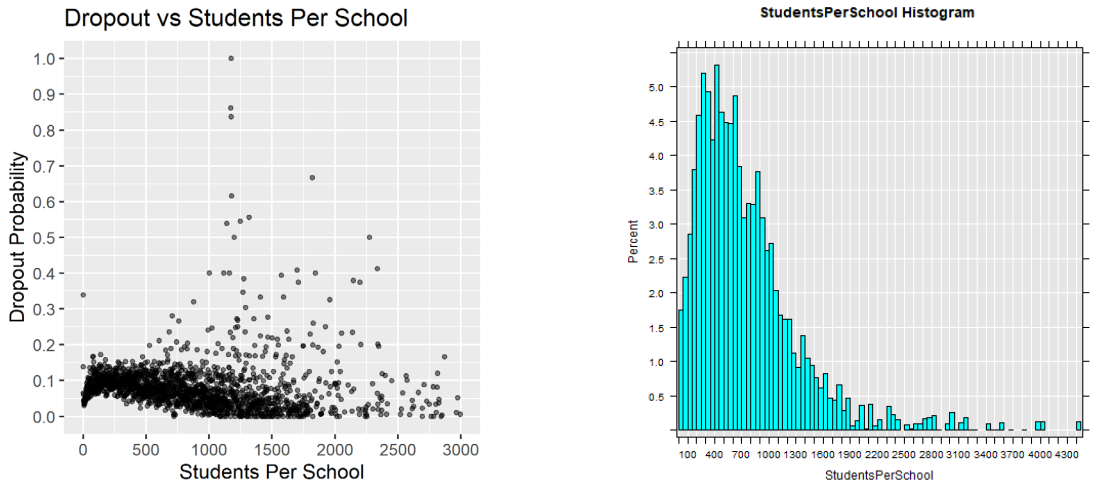


Table 11. Joint Correlation of all Variables

	DO	A	SM	SC	CC	SO	EO	SY	ETO	SYR	SPC	TH	TY	TS	SS	AG	ADG	GO	**STC**	**SCO**	**YSTC**	**YSC**
DropOut, DO	1,00	-0,23	-0,33	0,16	0,00	0,25	0,15	0,17	0,15	0,35	-0,03	-0,03	0,02	0,01	-0,03	0,22	0,31	0,04	0,05	0,06	0,00	0,00
Attendance, A	-0,23	1,00	0,33	-0,04	0,00	-0,24	-0,23	-0,18	-0,22	-0,27	-0,07	-0,01	-0,09	-0,09	-0,15	-0,20	-0,14	-0,01	-0,12	-0,13	-0,02	-0,07
SchoolMarks, SM	-0,33	0,33	1,00	-0,16	0,00	-0,32	-0,29	-0,39	-0,28	-0,51	-0,05	0,02	-0,04	-0,05	-0,01	-0,42	-0,27	-0,11	-0,07	-0,08	0,08	0,07
TotalSchoolChanges, SC	0,16	-0,04	-0,16	1,00	0,01	0,06	-0,13	0,05	-0,13	0,21	-0,06	-0,04	0,04	-0,01	-0,12	0,06	0,06	0,01	0,04	0,03	-0,04	-0,05
YearlyAverageClassChanges, CC	0,00	0,00	0,00	0,01	1,00	0,00	0,00	0,00	0,00	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
SchoolOrd, SO	0,25	-0,24	-0,32	0,06	0,00	1,00	0,35	0,29	0,31	0,17	-0,13	-0,05	0,08	0,05	-0,10	0,32	0,21	0,04	0,18	0,21	-0,02	-0,01
EducationLevelOrd, EO	0,15	-0,23	-0,29	-0,13	0,00	0,35	1,00	0,78	0,95	0,08	0,13	-0,06	-0,01	0,09	0,22	0,77	0,09	-0,01	-0,01	0,03	-0,02	0,03
SchoolYear, SY	0,17	-0,18	-0,39	0,05	0,00	0,29	0,78	1,00	0,78	0,10	0,12	-0,07	0,03	0,09	0,20	0,98	0,13	-0,01	-0,02	0,02	0,00	0,04
EducationTypeOrd, ETO	0,15	-0,22	-0,28	-0,13	0,00	0,31	0,95	0,78	1,00	0,08	0,12	-0,07	-0,01	0,08	0,22	0,77	0,09	0,01	-0,02	0,02	0,00	0,04
TotalSchoolYearRepeats, SYR	0,35	-0,27	-0,51	0,21	0,01	0,17	0,08	0,10	0,08	1,00	0,00	-0,01	0,03	0,03	0,00	0,13	0,16	0,07	0,04	0,05	-0,03	-0,02
StudentsPerClass, SPC	-0,03	-0,07	-0,05	-0,06	0,00	-0,13	0,13	0,12	0,12	0,00	1,00	0,09	-0,05	0,04	0,48	0,10	-0,09	-0,03	0,13	0,12	0,09	0,11
TeacherContractHours, TH	-0,03	-0,01	0,02	-0,04	0,00	-0,05	-0,06	-0,07	-0,07	-0,01	0,09	1,00	0,07	0,11	0,25	-0,07	-0,03	0,01	0,03	0,01	0,03	0,02
TeacherYearsService, TY	0,02	-0,09	-0,04	0,04	0,00	0,08	-0,01	0,03	-0,01	0,03	-0,05	0,07	1,00	0,80	0,00	0,03	0,02	0,00	-0,08	-0,06	-0,07	-0,05
TeacherYearsInSchool, TS	0,01	-0,09	-0,05	-0,01	0,00	0,05	0,09	0,09	0,08	0,03	0,04	0,11	0,80	1,00	0,11	0,09	0,01	0,00	-0,10	-0,07	-0,09	-0,07
StudentsPerSchool, SS	-0,03	-0,15	-0,01	-0,12	0,00	-0,10	0,22	0,20	0,22	0,00	0,48	0,25	0,00	0,11	1,00	0,18	-0,07	-0,01	0,05	0,06	0,19	0,22
Age, AG	0,22	-0,20	-0,42	0,06	0,00	0,32	0,77	0,98	0,77	0,13	0,10	-0,07	0,03	0,09	0,18	1,00	0,30	0,00	-0,01	0,02	0,01	0,04
AgeDifference_wr_LevelGrade, ADG	0,31	-0,14	-0,27	0,06	0,00	0,21	0,09	0,13	0,09	0,16	-0,09	-0,03	0,02	0,01	-0,07	0,30	1,00	0,07	0,01	0,01	0,02	0,01
GenderOrd,GO	0,04	-0,01	-0,11	0,01	0,00	0,04	-0,01	-0,01	0,01	0,07	-0,03	0,01	0,00	0,00	-0,01	0,00	0,07	1,00	0,00	0,00	0,00	-0,01
StudentCommuneOrd, STCO	0,05	-0,12	-0,07	0,04	0,00	0,18	-0,01	-0,02	-0,02	0,04	0,13	0,03	-0,08	-0,10	0,05	-0,01	0,01	0,00	1,00	0,80	-0,04	0,01
SchoolCommuneOrd, SCO	0,06	-0,13	-0,08	0,03	0,00	0,21	0,03	0,02	0,02	0,05	0,12	0,01	-0,06	-0,07	0,06	0,02	0,01	0,00	0,80	1,00	0,02	0,02
MeanIncome_StudentCommune, YSTC	0,00	-0,02	0,08	-0,04	0,00	-0,02	-0,02	0,00	0,00	-0,03	0,09	0,03	-0,07	-0,09	0,19	0,01	0,02	0,00	-0,04	0,02	1,00	0,89
MeanIncome_SchoolCommune, YSC	0,00	-0,07	0,07	-0,05	0,00	-0,01	0,03	0,04	0,04	-0,02	0,11	0,02	-0,05	-0,07	0,22	0,04	0,01	-0,01	0,01	0,02	0,89	1,00

Source: Government of Chile, Ministry of Education, 2018. http://datosabiertos.mineduc.cl/ and own elaboration.

Table 12. Individual Sensitivities

Rank	Variable	Individual sensitivity		
		Accuracy	Specificity	Sensitivity
1	TotalSchoolYearRepeats	91.86%	99.20%	10.85%
2	SchoolMarks	92.07%	99.50%	10.05%
3	AgeDifference_wr_LevelGrade	91.87%	99.32%	9.66%
4	Attendance	91.72%	99.92%	1.19%
5	TotalSchoolChanges	91.69%	100.00%	0.00%
6	YearlyAverageClassChanges	91.69%	100.00%	0.00%
7	EducationLevelOrd	91.69%	100.00%	0.00%
8	SchoolYear	91.69%	100.00%	0.00%
9	EducationTypeOrd	91.69%	100.00%	0.00%
10	StudentsPerClass	91.69%	100.00%	0.00%
11	TeacherContractHours	91.69%	100.00%	0.00%
12	TeacherYearsService	91.69%	100.00%	0.00%
13	TeacherYearsInSchool	91.69%	100.00%	0.00%
14	StudentsPerSchool	91.69%	100.00%	0.00%
15	Age	91.69%	100.00%	0.00%
16	GenderOrd	91.69%	100.00%	0.00%
17	MeanIncome_StudentCommune	91.69%	100.00%	0.00%
18	MeanIncome_SchoolCommune	91.69%	100.00%	0.00%

Source: own elaboration.

Table 13. Cumulative Sensitivities

Rank	Variable	Cumulative sensitivity					
		Accuracy	Specificity	Sensitivity	Δaccuracy	Δspecificity	Δsensitivity
1	TotalSchoolYearRepeats	91.86%	99.20%	10.85%			
2	SchoolMarks	92.10%	99.21%	13.69%	0.25%	0.01%	2.84%
3	AgeDifference_wr_LevelGrade	92.87%	98.90%	26.32%	0.76%	-0.31%	12.64%
4	Attendance	92.88%	98.76%	27.94%	0.01%	-0.13%	1.62%
5	TotalSchoolChanges	93.00%	98.84%	28.47%	0.12%	0.08%	0.52%
6	YearlyAverageClassChanges	93.00%	98.84%	28.47%	0.00%	0.00%	0.00%
7	EducationLevelOrd	93.67%	98.68%	38.28%	0.67%	-0.16%	9.81%
8	SchoolYear	93.85%	98.80%	39.22%	0.18%	0.11%	0.94%
9	EducationTypeOrd	93.85%	98.80%	39.22%	0.00%	0.00%	0.00%
10	StudentsPerClass	93.85%	98.76%	39.71%	0.01%	-0.04%	0.49%
11	TeacherContractHours	93.85%	98.76%	39.71%	0.00%	0.00%	0.00%
12	TeacherYearsService	93.85%	98.76%	39.71%	0.00%	0.00%	0.00%
13	TeacherYearsInSchool	93.85%	98.76%	39.71%	0.00%	0.00%	0.00%
14	StudentsPerSchool	93.84%	98.64%	40.91%	-0.01%	-0.12%	1.20%
15	Age	93.92%	98.58%	42.44%	0.07%	-0.06%	1.54%
16	GenderOrd	93.92%	98.58%	42.44%	0.00%	0.00%	0.00%
17	MeanIncome_StudentCommune	93.92%	98.60%	42.28%	0.00%	0.02%	-0.16%
18	MeanIncome_SchoolCommune	93.92%	98.60%	42.27%	0.00%	0.00%	-0.01%

Source: own elaboration.

Comparison of Accuracy, Specificity and Sensitivity of Algorithms

Finally, we present our results comparing each algorithm in terms of their respective accuracies, specificities and sensitivities, as outlined in the Table 14.

Table 14. Comparison of Algorithms

Algorithm	Training Time (hh:mm:ss)	Accuracy	Specificity	Sensitivity
Logit	0:00:49	92.97%	98.96%	28.65%
Decision Tree	0:06:58	93.71%	98.63%	40.94%
Random Forest	0:16:17	93.75%	98.82%	39.28%
Neural Network	0:17:22	93.80%	98.72%	41.05%

Source: own elaboration.

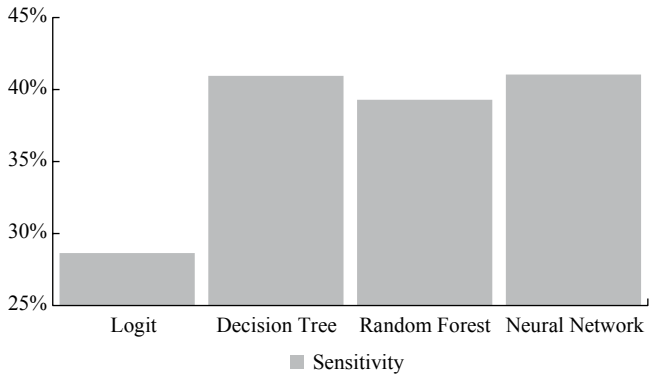
The four models have almost identical accuracies and specificities, approximately 93% and 98% respectively. The neural network is the most accurate and the logit model has the highest specificity; however, the differences are not significant.

The significant differences lie in the sensitivities, with the neural network having the greatest at 41%, 1.4 times that of the lowest: the logit model at 29%.

It is also interesting to note that there is a trade-off in terms of computing time: the neural network takes over 17 minutes to train, versus 49 seconds for the logit model.

The Figure 11 summarises these results.

Figure 11. Results by Algorithm (Comparison of Model Sensitivity)



In the graph, we can clearly observe that the sensitivity of the three machine learning algorithms is significantly greater than the traditional logit regression.

Conclusions and recommendations

We recall our research question: does machine learning enable more accurate early warning of school dropout in Chile? In Figure 11 we observe that the true positive rate

of the machine learning algorithms is significantly greater than that of logit regression, in other words, a greater percentage of real dropout cases are detected by these algorithms.

Before concluding the superiority of the machine learning algorithms, we must rule out the possibility that the higher accuracy is due solely to overfitting, i.e., that the models predict well on the same data with which they were trained but would perform poorly on the realistic scenario of a new data set. In effect, we can rule out overfitting because all four models were tested on a different data set than the training one (see the method section).

In addition, it is important to note that the predictions are made *in the year before* actual dropout takes place. This is an essential requisite for an *early* warning system, which must provide sufficient time for rescue interventions on the children at risk to be carried out.

We can therefore conclude that machine learning yields greater accuracy, in particular a greater true positive rate in predicting school dropout, and furthermore, it has been demonstrated with the data that is available in Chile, thus paving the way for continuing research and application in that country.

We acknowledge a limitation to our study. Binary models (where the dependent variable is binary: True/False) have as output a probability, and in selecting a probability threshold to define the prediction, there is a trade-off between the sensitivity and specificity, as illustrated in the ROC (Receiver Operating Characteristic) curve. In order to achieve an optimum model, it is important to find the probability threshold that maximises a metric (such as a Cobb-Douglas utility function) that takes into account both measures of accuracy. We did not do this, but simply chose the standard 0.5 probability threshold. The reason for this omission is that the purpose of our study was simply to demonstrate a "proof of concept" in the application of machine learning, and not to achieve an optimum model. We hope that other researchers and policy makers will "receive the baton" and perfect a model to be applied in practice in Chile.

We conclude with some recommendations for the application and implementation of practical early warning school dropout prediction systems that we hope will be useful, as well as possible research implications.

Optimise the model with a probability threshold that maximises a utility function of both sensitivity and specificity, for example:

$$U(x, y) = xy$$

where x = sensitivity and y = specificity.

Implement the neural network model, or alternatively the decision tree, with a computer application that provides a report of dropout probability for each student. This may be profitably used by the counsellors and social workers of each school.

Collect data of other variables that may be correlated with dropout, such as indicators of well-being, happiness and living standards. Analyse the correlation of these new variables with dropout.

This study was limited to prediction, and hence does not imply causality between the variables. We encourage other researchers to extend it to a study of causality, i.e., the root causes of school dropout, which would have important implications for public policy aimed at prevention.

In particular, it would be important to study in greater depth the relationship between the number of students per class and the dropout probability, since the scatterplot seems counterintuitive. This would have important implications for public policy in education.

We encourage other researchers to apply data science in multiple fields, and thus generate a myriad of benefits to society, such as more accurate medical diagnoses, efficient clean energy management, and improved macroeconomic forecasting.

References

Adelman, M., & Székely, M. (2017). An overview of school dropout in Central America: Unresolved Issues and new challenges for education progress. *European Journal of Educational Research*, *6*(3), 235-259. https://doi.org/10.12973/eu-jer.6.3.235

Alban, M. S., & Sánchez, D. M. (2018). Prediction of university dropout through technological factors: A case study in Ecuador. *Espacios*, *39*(52), 1-8.

Barros, T. M., Souza Neto, P. A., Silva, I., & Guedes, L. A. (2019). Predictive models for imbalanced data: A school dropout perspective. *Education Sciences*, *9*(4). https://doi.org/10.3390/educsci9040275

Dougherty, C. (2011). *Introduction to econometrics* (4th ed.). Oxford University Press.

Economist Intelligence Unit. (2015). *Big data evolution: Forging new corporate capabilities for the long term*. https://www.sas.com/en_us/offers/15q3/big-data-evolution.html (retrieved 2016)

Einav, L., & Levin, J. (2014). Economics in the age of big data. *Science* (346, 6210). https://doi.org/10.1126/science.1243089

Espíndola, E., & León, A. (2002). La deserción escolar en América Latina: un tema prioritario para la agenda

regional. *Revista Ibero-Americana de Educación*, *30*, 39-62.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An introduction to statistical learning: With applications in R* (2nd ed.). Springer Publishing Company.

Kleinbaum, D. G., Klein, M., & Pryor, E. R. (2010). *Logistic regression: A self-learning text* (3rd ed.). Springer Publishing Company.

Lee, S., & Chung, J. Y. (2019). The machine learning-based dropout early warning system for improving the performance of dropout prediction. *Applied Sciences (Switzerland)*, *9*(15). https://doi.org/10.3390/app9153093

Martinho, V. R. C., Nunes, C., & Minussi, C. R. (2013). An intelligent system for prediction of school dropout risk group in higher education classroom based on artificial neural networks. *Proceedings - International Conference on Tools with Artificial Intelligence.* ICTAI, May 2016, 159-166. https://doi.org/10.1109/ICTAI.2013.33

Mduma, & Neema. (2019). A survey of machine learning approaches and techniques for student dropout prediction. *Data Science Journal*, *53*(3). https://doi.org/https://doi.org/10.5334/dsj-2019-014

Melis, F., Díaz, R., & Palma, A. (2005). *Adolescentes y jóvenes que abandonan sus estudios antes de finalizar la enseñanza media: Principales tendencias*. https://docplayer.es/5865798-Adolescentes-y-jovenes-que-abandonan-sus-estudios-antes-de-finalizar-la-ensenanza-media-principales-tendencias-division-social-mideplan.html

Moshiri, S. and Cameron, N. (2000) Neural network versus econometric models in forecasting. *Journal of Forecasting*, 19, https://doi.org/10.1002/(SICI)1099-131X(200004)19:3%3C201::AID-FOR753%3E3.0.CO;2-4

Mineduc. (2014). *Programa de Transferencia de Convivencia Escolar y Alerta Temprana de la Deserción Escolar en Establecimientos Educacionales Municipales de la Región Metropolitana*. Unpublished document, Ministerio de Educación, Santiago, Chile.

Mullainathan and Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives, 31*(2), 2017, 87-106. http://doi.org/10.1257/jep.31.2.87

Municipalidad de Peñalolén. (2012). *Sistema de alerta temprana , un camino de prevención a la deserción escolar*. https://vinculacion.unab.cl/wp-content/uploads/2013/12/Presentación-Seminario-Trabajo-Infantil-Mun.-Peñalolén.pdf

Poggio, T., & Smale, S. (2003). The mathematics of learning: Dealing with data. *Notices of the American Mathematical Society (AMS), 50*(5), 537-544.

Secretaría de Educación Pública. (2011). *Sistema de alerta temprana (SIAT): lineamientos de operación de acciones de intervención*. https://transparencia.info.jalisco.gob.mx/sites/default/files/siat.pdf

Varian, H. (2014). Big Data: New tricks for econometrics tools to manipulate Big Data. *Journal of Economic Perspectives*, *28*(2), 3-28. http://dx.doi.org/10.1257/jep.28.2.3